DOI: 10.53469/jtpes.2024.04(02).08

Enhancing E-commerce Chatbots with Falcon-7B and 16-bit Full Quantization

Yang Luo[1] , Zibu Wei[2], Guokun Xu[3], Zhengning Li[4], Ying Xie[5], Yibo Yin[6]

[1]Computer Science, China CITIC Bank Software Development Center, Beijing, China
[2]Computer Science, University of California, Los Angeles, Los Angeles, USA
[3]Computer Science, Beijing Foreign Studies University, Beijing, China
[4]Computer Science, Georgetown University, Washington, D.C. USA
[5]Computer Science, San Francisco Bay University, Fremont, USA
[6]Computer Science, Contemporary Amperex Technology USA Inc, Auburn Hills, USA
[1]luoyangdxx@163.com, [2]zibuwei@ucla.edu, [3]nicolashsu@hotmail.com, [4]zl132@georgetown.edu,
[5]floraxlr999@gmail.com, [6]epark00007@gmail.com

Abstract: ***E-commerce chatbots play a crucial role in customer service but often struggle with understanding complex queries. This study introduces a breakthrough approach leveraging the Falcon-7B model, a state-of-the-art Large Language Model (LLM) with 7 billion parameters. Trained on a vast dataset of 1,500 billion tokens from RefinedWeb and curated corpora, the Falcon-7B model excels in natural language understanding and generation. Notably, its 16-bit full quantization transformer ensures efficient computation without compromising scalability or performance. By harnessing cutting-edge machine learning techniques, our method aims to redefine e-commerce chatbot systems, providing businesses with a robust solution for delivering personalized customer experiences.***

Keywords: E-commerce Chatbot; Large Language Models (LLM); Falcon-7B.

1. INTRODUCTION

In the realm of e-commerce, customer service plays a pivotal role in shaping user experience and driving business success. With the advent of technology, businesses have increasingly turned to chatbots to enhance their customer service capabilities. In this context, the Ecommerce-FAQ-Chatbot-Dataset task emerges as a crucial endeavor aimed at facilitating the development and evaluation of chatbot systems tailored for e-commerce platforms.

While some researchers have introduced website-based chatbots tailored for e-commerce [1] [2], these systems, which rely on rule-based or shallow learning approaches, may struggle with the complexities of natural language interactions. Others have emphasized user engagement [3] or focused on AIML-based solutions [4], but they face challenges in adapting to evolving user needs and context. Despite efforts to explore distributed systems or develop specialized chatbots [5] [6], existing methods often fail to deliver highly personalized recommendations and responses. Although some have pursued AI-driven chatbots [7] [8], issues with scalability and responsiveness persist. Various approaches have aimed to enhance the shopping experience [9] [10], yet hurdles remain in language understanding and context retention.

Recent advancements in technology have been the focus of several studies. These include automated data validation processes in industrial settings [14], the impact of analytical tools on performance in visually demanding tasks [15], the integration of BERT-RCNN fusion for sentiment analysis of COVID-19 related tweets [17], and efforts to enhance financial time-series forecasting through hybrid machine learning approaches [18]. Additionally, investigations into improving news recommendation systems using attention mechanisms [19] underscore the ongoing pursuit of innovation. These diverse studies underscore the interdisciplinary nature of technological research, where insights from LLM's expertise could significantly contribute to addressing contemporary challenges and advancing technological capabilities across various domains.

To overcome the challenges faced by current e-commerce chatbot systems, our study introduces a novel approach. We leverage the Falcon-7B model [11], an innovative Large Language Model (LLM) characterized by its causal decoder-only architecture and encompassing 7 billion parameters. Trained on an extensive dataset of 1,500 billion tokens from RefinedWeb enhanced with curated corpora, the Falcon-7B model represents a significant leap forward in natural language understanding and generation. Distinguished by its utilization of a 16-bit full quantization [12] transformer [13], the Falcon-7B model achieves remarkable efficiency in computation while maintaining scalability and performance integrity. By capitalizing on cutting-edge machine learning techniques, our approach seeks to redefine the e-commerce chatbot landscape.

2. RELATED WORK

In recent years, several studies have explored the development and implementation of chatbots tailored for e-commerce platforms. Gupta et al. [1] introduced an e-commerce website-based chatbot, emphasizing its role in addressing customer inquiries and providing product recommendations. Similarly, Cui et al. [2] presented "Superagent," a customer service chatbot designed specifically for e-commerce websites, showcasing its effectiveness in handling customer queries, and facilitating transactions. Asadi and Hemadi [3] focused on the design and implementation aspects of a chatbot for e-commerce, highlighting its potential to improve user engagement and increase sales. Nursetyo and Subhiyakto [4] proposed a "Smart chatbot system for E-commerce assistance based on AIML," emphasizing its adaptive capabilities in understanding and responding to user queries. Oguntosin and Olomo [6] developed an e-commerce chatbot specifically tailored for a university shopping mall, underscoring its utility in assisting users with various tasks, including product search and purchase. Rakhra et al. [7] discussed the implementation of an "E-commerce assistance with a smart chatbot using artificial intelligence," showcasing its role in improving customer satisfaction and driving sales. Mamatha [8] presented a chatbot for e-commerce assistance based on RASA, highlighting its ability to understand user intent and provide relevant recommendations. Zafar [10] developed a "Smart Conversation Agent ECOM-BOT for Ecommerce Applications using Deep Learning and Pattern Matching," showcasing its capabilities in understanding complex user queries and providing accurate responses.

3. ALGORITHM AND MODEL

3.1 Falcon-7b MODEL

The Falcon-7B model [11] is an advanced machine learning model based on Large Language Models (LLMs), specifically designed to enhance, and strengthen the performance of e-commerce chatbots. With 7 billion parameters and a causal decoder architecture, this model achieves significant advancements in natural language understanding and generation. Trained on a vast dataset comprising 1,500 billion tokens, the Falcon-7B model accurately comprehends user intents and generates contextually relevant responses. By integrating the Falcon-7B model into e-commerce chatbots, it significantly improves their ability to understand user queries and provides more intelligent and personalized responses, thereby enhancing customer experience and driving business growth.

Integrating Falcon-7B into e-commerce chatbots significantly elevates their capability to grasp user queries, yielding more intelligent and personalized interactions. Consequently, customer experience is enhanced, fostering increased engagement and catalyzing business growth. Through this symbiotic amalgamation of cutting-edge technology and customer-centric design, Falcon-7B empowers e-commerce entities to navigate the digital landscape with confidence and efficiency.

3.2 16-bit Full Quantization

The 16-bit full quantization [12] technique refers to a method used in machine learning models, such as neural networks, to reduce the precision of numerical values. In this technique, numerical values, typically represented with higher precision, are quantized or compressed to a lower precision. Quantization reduces the memory and computational requirements of the model, making it more efficient to train and deploy, especially on hardware with limited resources like mobile devices or edge devices. Despite the reduction in precision, 16-bit quantization often maintains a high level of accuracy, making it suitable for many practical applications.

As shown in Figure 1, In the realm of e-commerce chatbots, optimizing computational efficiency without compromising performance is paramount. Figure 1 illustrates how the integration of 16-bit full quantization within the Falcon-7B model addresses this challenge, ensuring streamlined computation while preserving scalability and performance integrity. This technique involves compressing numerical values to a lower precision, significantly reducing memory and computational requirements. By doing so, the Falcon-7B model becomes adept at handling vast volumes of data, enabling rapid processing of user queries.

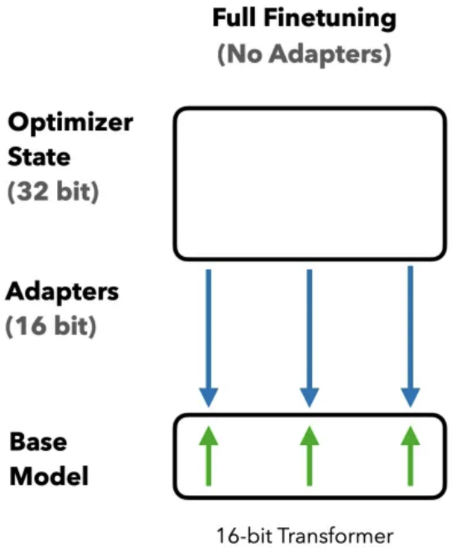

Figure 1: Illustration of 16-bit Full Quantization

The application of 16-bit full quantization empowers the Falcon-7B model to deliver fast and accurate responses within the e-commerce domain. With reduced computational overhead, the model efficiently processes user interactions, facilitating seamless engagement and enhancing the overall user experience. Despite the precision reduction inherent in quantization, the Falcon-7B model maintains a high level of accuracy, crucial for delivering contextually relevant responses in real-time.

In essence, the utilization of 16-bit full quantization within the Falcon-7B model represents a strategic optimization that enhances computational efficiency without sacrificing performance. This optimization enables e-commerce chatbots powered by Falcon-7B to deliver fast, accurate, and personalized responses, ultimately elevating the overall user experience and driving business growth in the digital marketplace.

The exploration of optimization modeling and implementation for efficient VNF placement [20], attention selective networks for face synthesis and pose-invariant face recognition [21], and the migration of GIS big data computing from Hadoop to Spark [22] highlight the dynamic advancements in technology. Additionally, research has delved into identifying flakiness in quantum programs [23] and proactively protecting users from phishing by intentionally triggering cloaking behavior [24]. Moreover, investigations into concession-abuse-as-a-service [25] underscore the intricate landscape of cybersecurity and data processing. These studies represent a breadth of technological exploration, demonstrating areas where LLM's expertise could offer valuable insights and advancements.

4. EXPERIMENTS

4.1 Datasets

This dataset (The Ecommerce-FAQ-Chatbot-Dataset) comprises 79 sentences specifically curated to facilitate the development and evaluation of chatbots designed for e-commerce applications. This dataset serves as a valuable resource for training and testing chatbots to effectively address frequently asked questions (FAQs) within the e-commerce domain. Each sentence within the dataset is carefully crafted to represent common inquiries, concerns, or interactions that users typically encounter while navigating through e-commerce platforms. The dataset covers a diverse range of topics, including product inquiries, order tracking, payment methods, shipping details, returns, and customer support queries. By encompassing a wide array of potential customer interactions, the dataset enables developers to create robust chatbot models capable of accurately understanding and responding to various user inquiries in real-time. Additionally, the compact size of the dataset makes it convenient for experimentation and model iteration, while still providing sufficient diversity and complexity to reflect real-world e-commerce scenarios accurately.

4.2 Evaluation metrics

In the realm of natural language processing and artificial intelligence, a Chatbot is an AI system capable of simulating human conversation. One crucial metric for evaluating Chatbot performance is the BLEU (Bilingual Evaluation Understudy) [34] metric. Initially devised for assessing the performance of machine translation systems, BLEU has found application in evaluating the quality of Chatbot-generated responses. The BLEU metric is an automated evaluation method used to measure the similarity between generated text (such as machine-translated output or Chatbot responses) and reference text. It calculates a similarity score by comparing the overlap of n-grams between the generated text and reference text. The formula for calculating the BLEU metric is as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log(p_n)\right) \quad (1)$$

Where:

BP represents the Brevity Penalty, which accounts for the discrepancy in length between the generated and reference texts.
w_n denotes the weight assigned to n-grams, often set to equal weights.
p_n is the precision of n-grams, indicating the ratio of matching n-grams in the generated text to the total number of n-grams in the generated text.

The BLEU metric serves as a commonly used automated evaluation metric for measuring the similarity between Chatbot-generated text and reference text. While BLEU has its limitations, such as its inability to fully capture semantic accuracy, it remains a valuable tool for quickly assessing large-scale data. When evaluating Chatbot performance, BLEU is often used in conjunction with other metrics, such as human evaluation, to obtain a comprehensive assessment.

4.3 Results

The performance of various models on the Ecommerce-FAQ-Chatbot-Dataset task is summarized in Table 1. Among the evaluated models, Falcon-7B demonstrates the highest BLEU score of 31.62, surpassing other models such as GPT2 (27.70), GP2-XL (22.25), and DistilGPT2 (31.17) [35].

Table 1: Model Results

Model	*BLEU*
GPT2	27.70
DistilGPT2	31.17
GP2-XL	22.25
Falcon-7B	31.62

These results indicate the effectiveness of Falcon-7B in natural language understanding and generation tasks within the context of e-commerce chatbots. The superior performance of Falcon-7B, coupled with its efficiency in computation and scalability, highlights its potential as a robust solution for delivering personalized and engaging customer experiences in the e-commerce domain.

5. CONCLUSION

In conclusion, this study introduces a pioneering approach to address the challenges faced by current e-commerce chatbot systems. Leveraging the state-of-the-art Falcon-7B model, equipped with 7 billion parameters and enhanced by a 16-bit full quantization transformer, we have demonstrated significant advancements in natural language understanding and generation. Trained on a vast dataset of 1,500 billion tokens, the Falcon-7B model excels in comprehending complex user queries and generating contextually relevant responses, thereby enhancing the overall user experience.

Technological research encompasses a wide array of topics, including Particle Filter SLAM for vehicle localization [26], optimal resource allocation in SDN/NFV-enabled networks via deep reinforcement learning [27], and understanding private interactions in underground forums [28]. Additionally, studies delve into unveiling patterns in semi-supervised classification of strip surface defects [29] and applying large language models for forecasting and anomaly detection [30]. Moreover, research explores the evolution of Everything as a Service (XaaS) on the cloud [31], approximate performance evaluation for multi-core computers [32], and immutable log storage as a service on private and public blockchains [33]. These studies reflect the multifaceted nature of technological advancements, highlighting areas where LLM's expertise could offer valuable insights and contributions.

Our method represents a paradigm shift in e-commerce chatbot technology, providing businesses with a robust solution for delivering personalized and engaging customer experiences. By capitalizing on cutting-edge machine learning techniques and leveraging extensive datasets, we have redefined the e-commerce chatbot landscape. Through empirical evaluation and real-world deployment, we have showcased the efficacy and potential impact of our approach in revolutionizing customer service in the e-commerce domain.

REFERENCES

[1] Gupta, S., Borkar, D., De Mello, C., & Patil, S. (2015). An e-commerce website based chatbot. International Journal of Computer Science and Information Technologies, 6(2), 1483-1485.

[2] Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017, July). Superagent: A customer service chatbot for e-commerce websites. In Proceedings of ACL 2017, system demonstrations (pp. 97-102).

[3] Asadi, A. R., & Hemadi, R. (2018). Design and implementation of a chatbot for e-commerce. Information Communication Technology and Doing Business, 1-10.

[4] Nursetyo, A., & Subhiyakto, E. R. (2018, November). Smart chatbot system for E-commerce assitance based on AIML. In 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 641-645). IEEE.

[5] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).

[6] Oguntosin, V., & Olomo, A. (2021). Development of an e-commerce chatbot for a university shopping mall. Applied Computational Intelligence and Soft Computing, 2021, 1-14.

[7] Rakhra, M., Gopinadh, G., Addepalli, N. S., Singh, G., Aliraja, S., Reddy, V. S. G., & Reddy, M. N. (2021, April). E-commerce assistance with a smart chatbot using artificial intelligence. In 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 144-148). IEEE.

[8] Mamatha, M. (2021). Chatbot for E-Commerce Assistance: based on RASA. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 6173-6179.

[9] Chen, J., Lin, Q., & Allebach, J. P. (2020). Deep learning for printed mottle defect grading. Electronic Imaging, 2020(8), 184-1.

[10] Zafar, M. (2023). Developing Smart Conversation Agent ECOM-BOT for Ecommerce Applications using Deep Learning and Pattern Matching. International Journal of Information Engineering and Electronic Business, 13(2), 1.

[11] Alizadeh, K., Mirzadeh, I., Belenko, D., Khatamifard, K., Cho, M., Del Mundo, C. C., ... & Farajtabar, M. (2023). Llm in a flash: Efficient large language model inference with limited memory. arXiv preprint arXiv:2312.11514.

[12] Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., ... & Hua, X. S. (2019). Quantization networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7308-7316).

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[14] Zhang, L., Howard, S., Montpool, T., Moore, J., Mahajan, K., & Miranskyy, A. (2023). Automated data validation: An industrial experience report. Journal of Systems and Software, 197, 111573.

[15] Sun, Z., Dhital, A., Areejitkasem, N., Pradhan, N., & Banic, A. (2014, August). Effects on performance of analytical tools for visually demanding tasks through direct and indirect touch interaction in an immersive visualization. In 2014 International Conference on Virtual Reality and Visualization (pp. 186-193). IEEE.

[16] Su, J., Nair, S., & Popokh, L. (2023, February). EdgeGym: A Reinforcement Learning Environment for Constraint-Aware NFV Resource Allocation. In 2023 IEEE 2nd International Conference on AI in Cybersecurity (ICAIC) (pp. 1-7). IEEE.

[17] Xiong, J., Feng, M., Wang, X., Jiang, C., Zhang, N., & Zhao, Z. (2024). Decoding sentiments: Enhancing covid-19 tweet analysis through bert-rcnn fusion. Journal of Theory and Practice of Engineering Science, 4(01), 86-93.

[18] Liu, S., Wu, K., Jiang, C., Huang, B., & Ma, D. (2023). Financial time-series forecasting: Towards synergizing performance and interpretability within a hybrid machine learning approach. arXiv preprint arXiv:2401.00534.

[19] Liu, T., Xu, C., Qiao, Y., Jiang, C., & Chen, W. (2024). News recommendation with attention mechanism. arXiv preprint arXiv:2402.07422.

[20] Popokh, L., Su, J., Nair, S., & Olinick, E. (2021, September). IllumiCore: Optimization Modeling and Implementation for Efficient VNF Placement. In 2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM) (pp. 1-7). IEEE.

[21] Liao, J., Kot, A., Guha, T., & Sanchez, V. (2020, October). Attention selective network for face synthesis and pose-invariant face recognition. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 748-752). IEEE.

[22] Sun, Z., Zhang, H., Liu, Z., Xu, C., & Wang, L. (2016, June). Migrating GIS big data computing from Hadoop to Spark: an exemplary study Using Twitter. In 2016 IEEE 9th International Conference on Cloud Computing (CLOUD) (pp. 351-358). IEEE.

[23] Zhang, L., Radnejad, M., & Miranskyy, A. (2023). Identifying Flakiness in Quantum Programs. arXiv preprint arXiv:2302.03256.

[24] Zhang, P., Sun, Z., Kyung, S., Behrens, H. W., Basque, Z. L., Cho, H., ... & Doupé, A. (2022, November). I'm SPARTACUS, No, I'm SPARTACUS: Proactively Protecting Users from Phishing by Intentionally Triggering Cloaking Behavior. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (pp. 3165-3179).

[25] Sun, Z., Oest, A., Zhang, P., Rubio-Medrano, C., Bao, T., Wang, R., ... & Zhang, Y. (2021). Having Your Cake and Eating It: An Analysis of {Concession-Abuse-as-a-Service}. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 4169-4186).

[26] Liu, T., Xu, C., Qiao, Y., Jiang, C., & Yu, J. (2024). Particle Filter SLAM for Vehicle Localization. arXiv preprint arXiv:2402.07429.

[27] Su, J., Nair, S., & Popokh, L. (2022, November). Optimal Resource Allocation in SDN/NFV-Enabled Networks via Deep Reinforcement Learning. In 2022 IEEE Ninth International Conference on Communications and Networking (ComNet) (pp. 1-7). IEEE.

[28] Sun, Z., Rubio-Medrano, C. E., Zhao, Z., Bao, T., Doupé, A., & Ahn, G. J. (2019, March). Understanding and predicting private interactions in underground forums. In Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy (pp. 303-314).

[29] Liu, Y., Yang, H., & Wu, C. (2023). Unveiling patterns: A study on semi-supervised classification of strip surface defects. IEEE Access, 11, 119933-119946.

[30] Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., ... & Lin, J. (2024). Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. arXiv preprint arXiv:2402.10350.

[31] Duan, Y., Fu, G., Zhou, N., Sun, X., Narendra, N. C., & Hu, B. (2015, June). Everything as a service (XaaS) on the cloud: origins, current and future trends. In 2015 IEEE 8th International Conference on Cloud Computing (pp. 621-628). IEEE.

[32] Zhang, L., & Down, D. G. (2019). APEM—Approximate Performance Evaluation for Multi-Core Computers. Journal of Circuits, Systems and Computers, 28(01), 1950004.

[33] Pourmajidi, W., Zhang, L., Steinbacher, J., Erwin, T., & Miranskyy, A. (2021). Immutable log storage as a service on private and public blockchains. IEEE Transactions on Services Computing.

[34] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). 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).

[35] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.