Humanizing Answers for Compatibility Questions in E-commerce using Large Language Models

André Gomes Regino[1,3], **Victor Hochgreb**[2], **Julio Cesar dos Reis**[1]

[1] Institute of Computing, University of Campinas – SP – Brazil

[2]GoBots

[3]Center for Information Technology Renato Archer –SP – Brazil

{andre.regino, jreis}@ic.unicamp.br,victor@gobots.com.br

Abstract. *Customer experience is a critical aspect of online purchase decisions. The service, the attendant's response, and how the customer is treated contribute to customer satisfaction. This article investigates using large language models for humanizing customer support in e-commerce. In particular, we address compatibility questions. Leveraging the infrastructure and dataset from an AI Brazilian startup, we compare the effectiveness of three different models to generate natural language answers in Portuguese. We generate human-like answers and evaluate them based on compatibility correctness, number of tokens, legibility, human likeness, and effect on the purchase. Our results highlight the effectiveness and drawbacks of the explored models in different temperature settings. This study improves customer experiences and provides guidance for e-commerce platforms in implementing humanized responses.*

1. Introduction

E-commerce adoption has experienced significant growth in recent years, revolutionizing how people shop and interact with products. With the convenience of online shopping, customers increasingly rely on product compatibility information to make purchasing decisions. In particular, when it comes to automotive products, customers often seek clarification on whether a specific product is compatible with their vehicle. This information is crucial to ensure customer satisfaction and avoid costly returns or compatibility issues. In general, the faster and more accurate the response, the greater the chances of purchasing because the customer is more confident in buying the compatible product from a seller who responded immediately, reducing the margin for competition.

We define a compatibility question as: given an item sold by a seller (*e.g.*, a XYZ brand tire), the product is considered compatible with a consumer's item (e.g., a 2021 Fiat Strada) if the sold product is suitable for the consumer's item. In our example, the XYZ brand tire must be compatible with the specified brand, model, and year of the car.

Traditionally, e-commerce platforms have relied on seller or attendant answers to address customer compatibility questions. If the seller has many products for sale and receives many compatibility questions, the response scale is compromised, which may lead to decreased sales. In addition, these answers often lack a human touch, appearing mechanical because the seller may have answered them many times. This impersonal approach can lead to customer frustration, reduced trust, and lower customer satisfaction [Tsai and Chuan 2023].

This problem extends to solutions that provide automatic responses. In our recent investigation [Regino et al. 2023], we illustrated how it is possible to use knowledge graphs to organize knowledge about automobile compatibilities in e-commerce and recommend products. The response given by the system that uses the knowledge graph is instantaneous, solving one of the problems of using human assistance in this context. However, the lack of humanization from the answers given still prevails. The answer follows the pattern: The product {product_name} is compatible/incompatible with the car {model_brand_year_of_car}.

Humanization refers to incorporating characteristics intrinsic to human beings, differentiating them from other animals and traditional automated responses. This involves creating interactions that are more natural, understanding, and compassionate [Legrand et al. 1991, dos Santos Viriato et al. 2023].

To bridge this gap and enhance customers' experience, integrating Large Language Models (LLMs) into customer support systems has emerged as a promising solution. Large language models, such as GPT 3.5 from OpenAI [Brown et al. 2020] and Bloom [Scao et al. 2022], have demonstrated remarkable capabilities in understanding and generating human-like text. LLMs have been adopted in e-commerce in several tasks, such as automatic product labeling [Chen et al. 2023], recommender systems [Lin et al. 2023], and others. Leveraging the infrastructure and dataset of GoBots, an AI Brazilian startup, we explore these models to humanize answers to customer compatibility questions in an e-commerce context. We explore the capabilities of LLMs in a Portuguese context once the clients from the Brazilian startup (where this study was conducted) are Portuguese native speakers.

In our investigation, we hypothesize that LLMs can improve the humanization of e-commerce compatibility questions written in Brazilian Portuguese. To the best of our knowledge, our study is the first effort to use LLMs to humanize answers in this context.

This study demonstrates how large language models can provide personalized and humanized answers to customer compatibility queries. By leveraging the dataset of an AI Brazilian startup with multiple e-commerce platforms as customers, we determine if these models can generate answers that improve customer satisfaction and engagement.

In this research, we employ a dataset of product names, customer questions, and seller-provided answers. We use a single prompt containing a description of the task and four examples in Portuguese to help the models generate the humanized answers. Additionally, we experiment with varying the randomness of the model (using temperature) to evaluate their impact on generating human-like answers. In our experimental analyses, we assess the effectiveness of Bloom [Scao et al. 2022], Qwen [**?**], and Mistral [Jiang et al. 2023] in automatically producing humanized answers for customer compatibility questions. Our findings contribute to the growing research on humanizing customer support in e-commerce. We discuss the potential of large language models to enhance customer experiences in a Portuguese context.

This article is organized as follows: Section 2 describes related work. Section 3 shows our methodology to humanize and evaluate the answers. Section 4 reports on our obtained results. Section 5 discusses our findings and open research challenges. Section 6 draws conclusion remarks.

2. Related Work

Sant'Anna *et al.* [Sant'Anna et al. 2020] proposed a knowledge graph to store questions and answers from Brazilian e-commerce stores. The answers come from attendants and are represented in an RDF triple format. Instead of using a human attendant to answer repeated questions, the triples from the knowledge graph are used as the source for automatic answer generation. Their work uses natural language processing tools to identify entities and intents from the questions and products.

Cheng *et al.* [Cheng et al. 2021] examined consumer trust in text-based chatbots in e-commerce. They found that chatbot attributes like empathy and friendliness boost consumer trust. Task complexity weakens the trust-building influence of friendliness but does not affect empathy. Additionally, their work revealed that greater consumer trust leads to increased reliance on the chatbot and reduced resistance, with the positive effect on reliance being stronger than the reduction in resistance.

Li and Wang [Li and Wang 2023] investigated the impact of language style in chatbot communication within e-commerce. They demonstrated that using informal language in chatbot interactions boosts customer engagement and influences brand attitude among existing customers. This may increase their intention to continue using the chatbot. Informal language can be detrimental for new or non-customers, as it fails to align with their expectations for more formal interactions. The study highlighted the need for brand managers to adjust chatbot language based on the customer's relationship with the brand to provide optimal user experiences and enhance customer retention.

Our originality lies in using language models to humanize responses in Brazilian e-commerce. We originally evaluate to which extent multiple models and temperatures handle the transformation of responses written in natural language into humanized responses, incorporating user engagement characteristics described in Li and Wang [Li and Wang 2023] and Cheng *et al.* [Cheng et al. 2021]. To the best of our knowledge, this study is the first effort to use LLMs to humanize answers in e-commerce.

3. Methodology

This section presents our method (Section 3.1) and evaluates it (Sections 3.2, 3.3, 3.4, 3.5 and 3.6).

3.1. Proposed Method

Our solution method involves generating humanized compatibility responses in e-commerce scenarios. Given a product, a "compatibility question" about that product, and a response from either a human attendant or an automated system, our solution generates a humanized response based on the input through a large language model.

3.2. Evaluation Overview

Figure 1 presents our methodology to evaluate the generated answers produced by our method. First, we randomly selected 100 products, questions, and answers from the automotive context in e-commerce. This dataset is composed of Portuguese products, questions, and answers. Section 3.3 describes our dataset in more detail. The second

component is the prompt. This component uses the few shots technique. It is characterized by the set of instructions and important examples provided to the model to generate the appropriate answers [Brown et al. 2020].

The prompt, temperature, and the third element describe the response expected by the model in terms of sentence structure, tone of speech, and level of randomness in the generated responses. These first three components are necessary for step A of the methodology: the parameter/input setting.

Afterward, each of the chosen models is executed (step B of Figure 1) with each of the lines of the dataset, each temperature, and prompt (Section 3.4). Such executions generate 600 humanized responses, responses that are evaluated (step C of Figure 1) through 5 different criteria, discussed in more detail in Section 3.5.

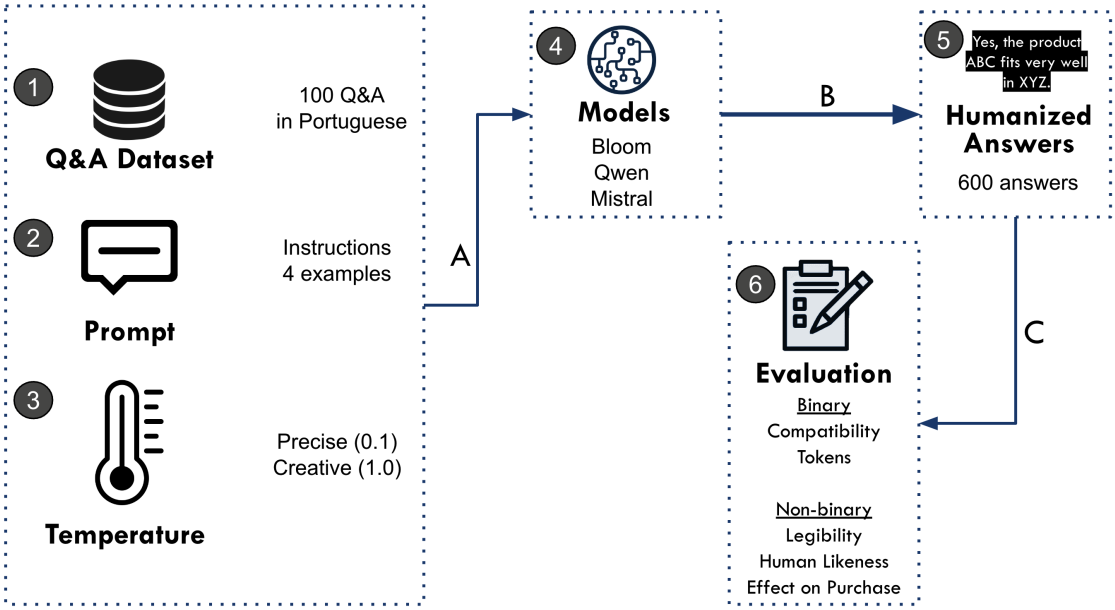

Figure 1. Our methodology to evaluate humanized answers. It comprises six components: 1 - the dataset, 2 - the prompt, 3- the temperatures, 4 - the models, 5 - the humanized answer, and 6 - the evaluation. The stages/actions to connect the components are represented by the letters A (Set parameters/input), B (Execute the model), and C (Evaluate results).

3.3. Dataset

We leverage the dataset that an AI Brazilian startup provides. This startup specializes in using AI to solve problems in e-commerce platforms. This startup has Latin American clients, allowing us to access a real-world dataset of customer compatibility questions and seller-provided answers. The dataset mainly comprises various automotive products, questions, answers, and purchases. This includes information in Portuguese, enabling us to address compatibility concerns across multiple Brazilian markets.

Table 1 presents two simplified dataset instance examples. According to the response, the first example shows a compatible product, whereas the second presents an incompatibility between the product and the customer's car. The response shown to the user can either be a response from a human attendant or an automatically built response.

The automatic answer only happens when the knowledge about the compatibility between the product and the car has already been stored, in case the answer has already been answered previously. The information about compatibilities is stored in a knowledge graph. More details on how this knowledge graph detects compatibility intents and the entities (car, brand, year) are described in our previous work [blind review].

For our experiments, we used 100 examples of customer compatibility questions and corresponding provided answers. Those examples were randomly selected from ten different stores to account for the diversity of data in the automobile context of the e-commerce platforms. This selection allows us to evaluate the effectiveness of the LLMs in generating human-like answers across different stores. These stores have different ways of responding to customers: they differ in the use of regionalisms, abbreviations and response length, among other characteristics. They sell various products with varying range of prices. Data was collected during the second semester of 2023.

Table 1. Two examples of the dataset instances used in the methodology.

Example #1	
Product	*Correia Poly V Alternador Fiesta*
Question	*Boa tarde! Serve no fiesta 2007, 1.0. 8valvulas, direção manual?*
Answer	*a peça anunciada é compatível*

Example #2	
Product	*Parachoque Dianteiro Ford Ka- 2003 A 2007- 3 Partes Completo*
Question	*Boa noite serar servi no ford ka 98?*
Answer	*Boa tarde! Tudo bem? Não é compatível.*

3.4. Models

To leverage the capabilities of LLMs in our context, we designed an appropriate input format for generating human-like answers. For LLMs, the standard input is the prompt. It is a strategy in which users of LLMs create natural language instructions or specifications, allowing them to shape the generated output or response of the models [Arora et al. 2022]. Instead of depending on explicit instructions or programming, prompts serve as inputs that directly influence the language model's behavior and output. The prompt we developed comprises two main parts: the instructions and the examples. The same prompt is used in all the experiments. The instruction part describes what we want the LLM to do and how. We clearly state that we want a humanized compatibility answer. The examples contain four examples, each including four main components: the product's name, the customer's compatibility question, the answer written by the seller/attendant or by the Knowledge Graph, and a humanized answer. We used four examples because we consider this number a balance between providing sufficient context to the model and maintaining prompt efficiency. The authors of this work chose these examples to cover a range of typical customer compatibility inquiries and responses, including compatibility and no-compatibility answers.

In our experimental evaluation, we used three LLMs to generate human-like answers: Bloom, Qwen, and Mistral. These models have demonstrated proficiency in natural language understanding and generation tasks. By feeding the structured prompt into

these models, we leverage their capacity to comprehend customer compatibility questions and produce personalized and coherent answers.

We chose Bloom, with 7 billion parameters, because it is open source and presents satisfactory results in Q&A tasks. Its training dataset included Portuguese texts. We chose to use the Qwen 1.5 model, with 7 billion parameters, because it is also open source; it was developed by the largest e-commerce company in the world, Alibaba, and consequently, it was created in an e-commerce environment, which is also the environment of our dataset. Its training dataset is multilingual and consists mainly of English and Chinese texts. We chose the Mistral model, which has 7 billion parameters, because it is open-source and offers lower latency than other models.

Temperature is a crucial parameter in controlling the randomness of the generated text. A higher temperature value produces more diverse and creative responses. A lower value produces more focused and deterministic answers. The range of temperature values allows us to investigate the trade-off between response creativity and adherence to customer expectations. Figure 1 presents two temperatures explored in our experiments: "more precise" (0.1), and "more creative" (1.0).

The models are combined with the temperatures in the following configurations: B01 (Bloom with temperature 0.1); B10 (Bloom with temperature 1.0); Q01 (Qwen with temperature 0.1) and Q10 (Qwen with temperature 1.0); M01 (Mistral with temperature 0.1) and M10 (Mistral with temperature 1.0).

The rationale behind choosing these temperature settings is that we aim to investigate whether responses from configurations with lower temperatures, meaning more straight, are more human-like, less human-like, or show no difference compared to responses generated by configurations with higher temperatures, which may produce more "human-like responses", but with more potential hallucinations and off-topic answers.

3.5. Evaluation Criteria

We describe the evaluation criteria employed to assess the quality of the generated answers. These five criteria were chosen because they comprehensively cover the responses' functional and qualitative aspects. They focus on the technical accuracy and the human-centered qualities of the responses. Other criteria were considered, but, to our knowledge, these were sufficient due to their ability to capture the key dimensions of response quality without redundancy. The criteria include:

- **Compatibility**: evaluates if the generated answer is right, stating that the product and the car are compatible. The compatibility is a binary-based criterion. It verifies if the answer has the correct compatibility/incompatibility statement. Before running the model on the dataset, the authors created a gold standard regarding compatibility, reading each answer and checking if states that the products are compatible or not. This allowed for the evaluation of how many humanized responses correctly identified compatibility or incompatibility;
- **Tokens**: the number of tokens produced by the model corresponds to a number that quantitatively describes the size of the answer. This criterion is important since it is the standard way to calculate the costs of LLM. 1000 tokens are approximately equal to 750 words;

- **Legibility**: assesses the readability and clarity of the generated answers. It considers factors such as sentence structure, punctuation, and readability. A legibility score of 1 (given by the evaluators) indicates answers that are impossible to read, while scores closer to 5 indicate legible answers with no or fewer grammar errors;
- **Human Likeness**: rates the degree to which the generated answers resemble human responses regarding tone, style, and overall naturalness. The human likeness score of 1 indicates short, meaningless, and less humanized answers, whereas score of 5 indicates answers that a human could provide in a humanized way;
- **Effect on Purchase**: rates how much the given automatically generated response could positively or negatively influence the purchase. The effect on purchase score ranges from 1, where, after reading the response, the evaluator would completely lose interest in the purchase, to 5, where the evaluator would be more inclined to make the purchase given the well-written response.

The final results are evaluated by considering the compatibility scores, the number of tokens, and the mean value of legibility, human likeness, and effect on purchase.

3.6. Evaluation Procedure

We used 45 Brazilian evaluators to conduct the evaluation. Native speakers must evaluate because the quality of humanized sentences is linked to aspects that native speakers can capture more easily. The evaluators were undergraduate students enrolled in courses that cover human-computer interaction elements and user experience, including humanization aspects. These students were chosen because their academic background and focus on user interface and experience make them well-suited to assess the nuances of humanized responses.

Each evaluator received a set of 13 products, along with customer questions and answers generated by the LLM, randomly selected from a dataset containing 600 product/question pairs. The evaluators were not informed that the responses were generated by an LLM, ensuring unbiased assessment. The evaluation process involved the following steps: 1) Reading each product, the corresponding customer question, and the answer provided by the LLM; 2) Evaluating the response based on three criteria: legibility, human likeness, and effect on purchase; 3) Assigning a score from 1 to 5 for each criterion for every response (see Table 2).

Table 2. Non-binary evaluation criteria

Score	Legibility	Human Likeness	Effect on Purchase
1	Impossible to read	Off-topic/many hallucinations	The response would make the buyer lose interest
2	Readable, but did not answer the question	Some hallucinations	The response generated distrust in the purchase
3	Many orthographic and semantic errors	Very short, direct, robotic text	The response is indifferent, it would not influence the purchase
4	Some orthographic and semantic errors	Humanized, but overly cordial	The response is good, but it would not influence the purchase
5	Readable and error-free sentence	Humanized	The response would generate interest in the purchase

Each evaluator assessed 3 criteria for 13 different responses, resulting in 39 evaluations per evaluator. This number was chosen to ensure the evaluation process was manageable, taking approximately ten minutes for each evaluator to complete. The scores given by each evaluator were averaged for each of the three criteria across the six configurations (B01, B10, Q01, Q10, M01, M10). This means we calculated an average legibility score for each configuration and average scores for human likeness and effect

on purchase. By comparing these average scores, we evaluated and compared the effectiveness of each configuration according to the three criteria.

4. Results

Table 3 presents the compatibility accuracy and number of tokens results achieved combining the three studied models and two temperatures (B01, B10, Q01, Q10, M01, M10). The percentages in the Compatibility column represent the accuracy rates achieved by the models in the respective techniques. These results were achieved after comparison with the previously generated gold standard. They were calculated by the ratio between the number of true positives (products correctly answered as compatible) and true negatives (products correctly answered as not compatible) over the total per configuration, which is 100. The number in the column Tokens refers to the number of tokens generated by the model in the humanized answer, the output. This value was calculated from the average token count of each response.

Table 3. Sum of results returned by each test set. The columns represent the criteria. The lines represent each test set: B01 stands for Bloom with temperature 0.1; B10 stands for Bloom with temperature 1.0; Q01 stands for Qwen with temperature 0.1 and Q10 stands for Qwen with temperature 1.0; M01 stands for Mistral with temperature 0.1 and M10 stands for Mistral with temperature 1.0. The compatibility results indicate how many humanized answers from the total (100) match the correct compatibility/incompatibility answer. The tokens result represents the average tokens per answer, followed by the minimum number and max number of tokens generated.

	Compatibility (%)	#Tokens
B01	91	18.88 (3 - 63)
B10	92	19.68 (3 - 63)
Q01	53	8.85 (1 - 89)
Q10	30	12.41 (1 - 59)
M01	95	21.07 (5 - 86)
M10	95	21.7 (5 - 89)

Figure 2 and Figure 3 show the quality of the humanized answers regarding the non-binary evaluations. The results comprise four graphs: legibility results, human-likeness results, effect on purchase results, and average results. In Figure 2 and in the first graph of Figure 3, the x-axis of each graph represents a different model-temperature configuration (B01, B10, M01, M10, Q01 and Q10). The y-axis describes how many humanized answers got each score (from 1 to 5) in the x-axis.

The first graph (number 1 in Figure 2) describes the legibility score. For example, the evaluators categorized 82 humanized answers from B01 with the maximum score (5) and 2 humanized answers from the Q01 configuration with the minimum score (1). The second graph (number 2 in Figure 2) describes the humanization score. For example, the evaluators categorized 3 humanized answers from B10 with score 2 and 34 humanized answer from Q10 configuration with score 3.

The third graph (number 3 in Figure 3) describes the effect on purchase score. For example, the evaluators categorized 44 humanized answers from Q10 with the minimum score (1) and 22 humanized answer from Q10 with the maximum score (5). The fourth graph (number 4 in Figure 3) synthetizes, using mean values, the results from the other three graphs. The first column of each configuration set (B01 to Q10) shows the mean

value of legibility score, based on the legibility graph of graph 1 (Figure 2). The second and third column of each configuration set shows the mean value of humanization and the effect on purchase score, respectively. The fourth column shows the mean value of the three evaluation criteria. For example, the M10 configuration set scored the highest (4.29), while Q10 scored the lowest (2.72).

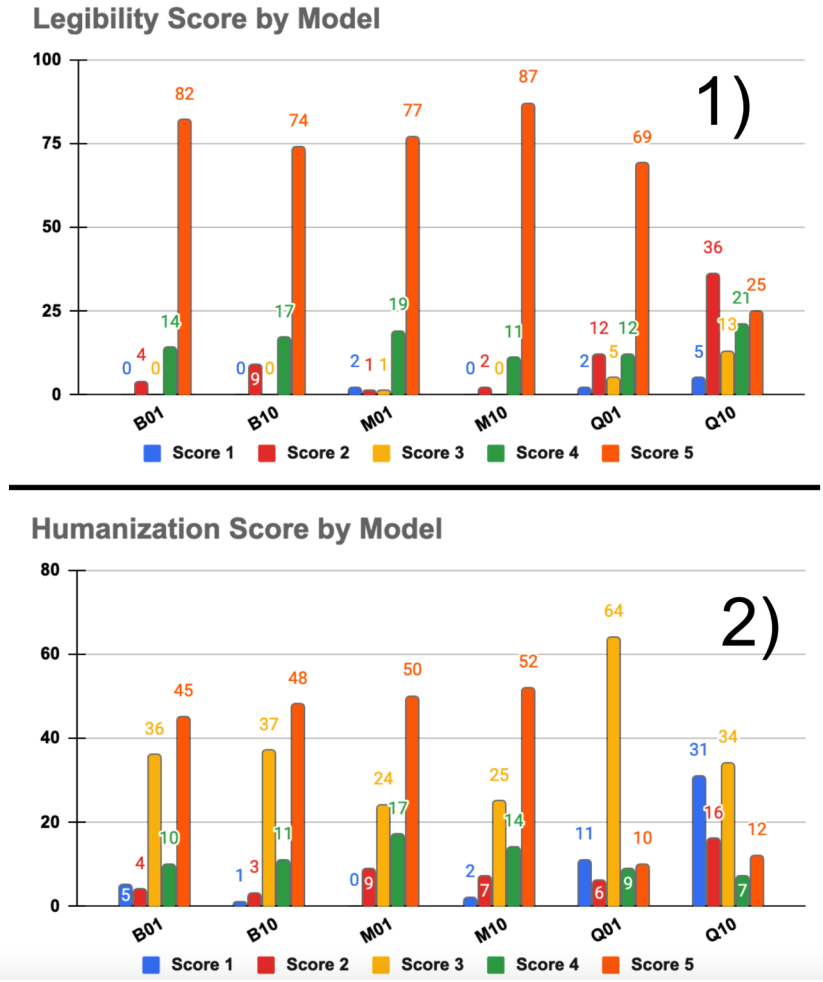

Figure 2. Results of legibility and human likeness. The first graph shows the readability results, and the second is the human likeness. There are 100 responses for each model-temperature configuration (represented in the x-axis), resulting in 600 responses evaluated. The colors represent the evaluators' scores (ranging from 1 to 5). The y-axis measures the count by the score for each configuration.

Compatibility analysis. Overall, the Bloom model performs consistently well across both temperature values regarding compatibility, with high percentages of correct results (91 and 92%). Mistral achieved the best results in terms of compatibility, with 95% accuracy in both temperatures (0.1 and 1.0). Qwen produced many wrong responses in terms of compatibility. Only half and a third of the responses with temperature 0.1 and 1.0 matched the expected compatibility output. We found several cases in which, instead of presenting false positives (responding that a product is compatible, but it is not) and false negatives (responding that a product is not compatible, but it is), Qwen did not answer the customer's question or hallucinated, answering other random questions. We observed that the performance of Qwen produced results analogous to the outcomes of a fair coin

toss, with a 50% chance of correctly predicting compatibility between the consumer item and the item sold by the retailer. For this reason, we do not recommend using this model, with the parameters defined in Section 3 (model size and temperature), to answer compatibility questions in Portuguese in the context of e-commerce. Even though the model generated humanized responses and encouraged purchases, providing incorrect compatibility information to the customer can negatively affect the post-sale process. This can lead to numerous complaints and customer dissatisfaction due to inaccurate information.

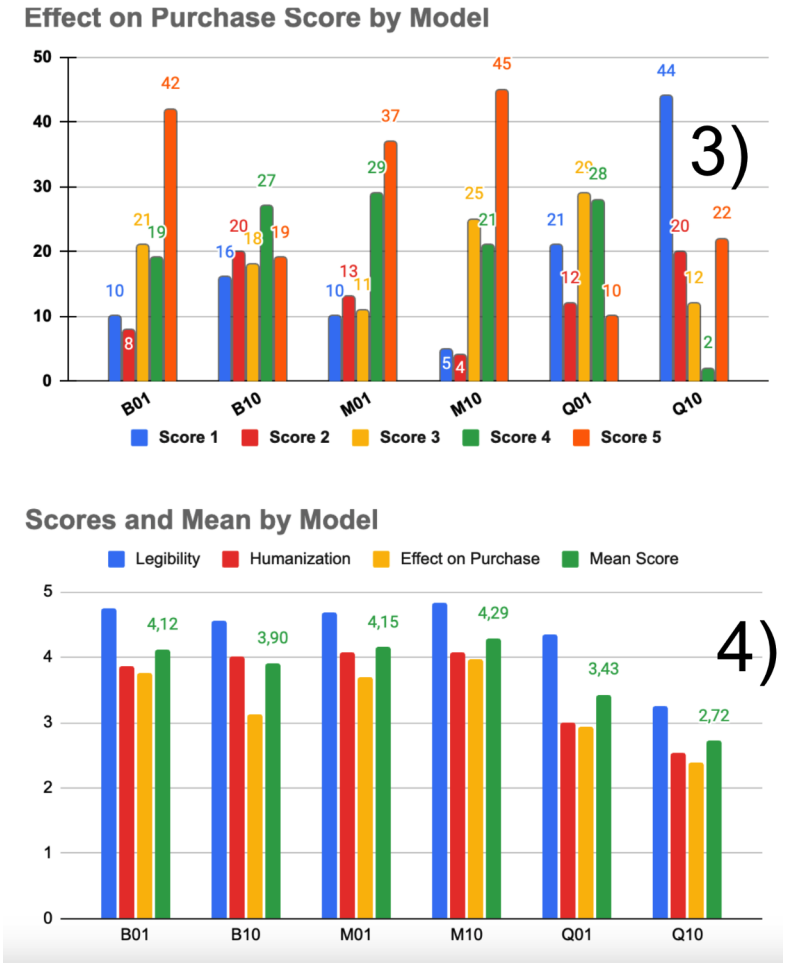

Figure 3. Results of effect on purchases and mean scores. For effect on the purchase graph, 100 responses for each model-temperature configuration (represented in the x-axis), resulting in 600 responses evaluated. The colors represent the evaluators' score (ranging from 1 to 5). The y-axis measures the count by score for each configuration. The second graph shows the mean value by each criteria: legibility, humanization and effect on purchase. The green column shows the mean of the three criteria.

Number of tokens analysis. The number of tokens used to generate the response by Bloom is slightly less than Mistral and higher than Qwen. This characteristic is relevant for systems in production, with a constant flow of responses. The lower the number of tokens, the lower the amount paid per response. However, the number of tokens and low cost do not always refer to the high quality of responses: Qwen, with Q01 and Q10, produced approximately half of the tokens than Bloom and Mistral, but a poor result

in compatibility match. In summary, fewer tokens do not necessarily imply humanized, assertive, and direct responses.

Legibility analysis. Based on the first graph in Figure 2, we observed that five of the six configurations achieved excellent results in terms of legibility. Specifically, 96%, 91%, 97%, 98%, and 81% of the answers from the B01, B10, M01, M10, and Q01 configurations received a score of 4 or higher for legibility. Overall, 85% of the 600 results (100 for each configuration) scored 4 or higher. This implies that the evaluators noted only a few grammatical (syntactic and semantic) errors in the humanized answers, which did not significantly compromise readability. The worst results came from the Qwen model, particularly the Q10 configuration. Only 46% of Qwen with temperature 1.0 results received a score of 4 or 5, making it an illegible and inappropriate configuration for use. In this configuration, 54% of the responses were either impossible to read, did not answer the question, or had many grammatical errors.

Human likeness analysis. For the human likeness evaluation criterion (graph 2 in Figure 2), the results were concentrated in scores 3 and 5. In summary, these results indicate incidences of short and automatic sentences (score 3), mainly in the tests carried out with the Qwen model, and fully humanized sentences (score 5), mainly in the tests with the Bloom and Mistral models. These two models produced similar humanized results, ranging from 45 to 52 responses scoring a 5. The difference between their results is that the Bloom model (B01 and B10) produced more score 3 results, while the Mistral model (M01 and M10) produced more score 4 results. This implies that if the intention is to produce shorter, more robotic, and direct answers, Bloom is suitable; for more humanized but overly cordial answers, Mistral is more adequate. Qwen models produced approximately 84% of all score 1 results, indicating many hallucinations in the answers.

Effect on purchase analysis. For the last criterion, the effect on purchase (graph 3 in Figure 3), the results demonstrate that the M10 configuration produced the majority of the best results (21% and 45% scoring 4 and 5, respectively) and the minority of the worst results (5% and 4% scoring 1 and 2, respectively). The Q10 configuration produced the worst results (44% and 20% scoring 1 and 2, respectively), while the B10 configuration produced similar results across all scores (ranging from 16% to 27%). In summary, out of the 600 responses, 50% were classified as good, indicating they would generate interest in the purchase (scores 4 and 5), while 30.5% of the responses would make the customer lose interest or distrust the purchase (scores 1 and 2).

Mean score analysis. Graph 4 in Figure 3 summarizes the results. The M10 configuration produced the best results for all non-binary criteria, with a mean score of 4.29. This promising result is due to the high scores achieved in each criterion, indicating that this configuration produced legible, humanized, convincing, and correctly compatible answers. In contrast, the Q10 configuration produced the worst results for all non-binary criteria, with a mean score of 2.72.

5. Discussion

Overall, our results demonstrated high accuracy rates in addressing compatibility concerns, with two of three models consistently providing relevant and informative responses. This includes evidence concerning our initial objective of using large language models to humanize customer support interactions and enhance the shopping experience.

The implications of using LLMs for humanizing customer support in e-commerce are significant. By leveraging these models, e-commerce platforms can provide personalized and contextually relevant answers to compatibility queries, fostering customer trust, satisfaction, and engagement. The human-like nature of the generated responses enhances the customer experience, reducing the perception of interacting with a computer-based system and increasing the perceived empathy and understanding from the seller's side.

The results suggest that a less conservative approach, not limiting randomness and focusing on conformity, may be preferred for generating accurate and reliable answers, once the configuration with temperature 1.0 achieved the best result (4.29). Conservative approaches (temperature 0.1) also achieved good results with B01 and M01 configurations. However, we note that the champion model selection may vary depending on the specific requirements of the e-commerce platform or the target customer segment.

To further improve and advance the field, future research can explore integrating rules-based approaches with large language models to enhance the humanization of answers. By incorporating heuristics that capture human-like decision-making processes or ethical guidelines, we can fine-tune the responses generated by the models, ensuring they align with desired standards of empathy, understanding, and ethical conduct.

6. Conclusion

Retaining customer attention and providing an excellent user experience through humanizing responses in e-commerce platforms is challenging. By providing more human-like answers, e-commerce platforms can foster customer trust, improve customer satisfaction, and drive business growth. Our study demonstrated the potential of LLMs to humanize customer support interactions in the e-commerce context. Our findings highlighted the effectiveness of Mistral and Bloom in generating human-like answers in Brazilian Portuguese, potentially improving customer satisfaction and driving engagement. In particular, the Mistral model achieved 95% accuracy in correctly answering compatibility questions by obtaining an average score of 4.29 out of 5 in evaluating criteria such as legibility, humanization of the response, and positive effect on purchase. Future research focuses on reducing LLMs hallucinations in the generated answers and exploring novel prompt engineering techniques, like Chain of Thoughts.

Acknowledgments

This study was financed by the National Council for Scientific and Technological Development - Brazil (CNPq) process number 140213/2021-0. In addition, this research was partially funded by the São Paulo Research Foundation (FAPESP) (grants #2022/13694-0, #2022/15816-5 and #2024/07716-6). The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

Arora, S., Narayan, A., Chen, M. F., Orr, L. J., Guha, N., Bhatia, K., Chami, I., Sala, F., and Ré, C. (2022). Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, J., Ma, L., Li, X., Thakurdesai, N., Xu, J., Cho, J. H., Nag, K., Korpeoglu, E., Kumar, S., and Achan, K. (2023). Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with llms. *arXiv preprint arXiv:2305.09858*.

Cheng, X., Bao, Y., Zarifis, A., Gong, W., and Mou, J. (2021). Exploring consumers' response to text-based chatbots in e-commerce: the moderating role of task complexity and chatbot disclosure. *Internet Research*, 32(2):496–517.

dos Santos Viriato, P. J., de Souza, R. R., Villas, L. A., and dos Reis, J. C. (2023). Revealing chatbot humanization impact factors. In Kurosu, M. and Hashizume, A., editors, *Human-Computer Interaction - Thematic Area, HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23-28, 2023, Proceedings, Part III*, volume 14013 of *Lecture Notes in Computer Science*, pages 294–313. Springer.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Legrand, G., Rodrigues, A., and Gama, J. (1991). *Dicionário de filosofia*.

Li, M. and Wang, R. (2023). Chatbots in e-commerce: The effect of chatbot language style on customers' continuance usage intention and attitude toward brand. *Journal of Retailing and Consumer Services*, 71:103209.

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Li, X., Zhu, C., Guo, H., Yu, Y., Tang, R., et al. (2023). How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817*.

Regino, A. G., Caus, R. O., Hochgreb, V., and Reis, J. C. d. (2023). Leveraging knowledge graphs for e-commerce product recommendations. *SN Computer Science*, 4(5):689.

Sant'Anna, D. T., Caus, R. O., dos Santos Ramos, L., Hochgreb, V., and dos Reis, J. C. (2020). Generating knowledge graphs from unstructured texts: Experiences in the e-commerce field for question answering. In *Advances in Semantics and Linked Data: Joint Workshop Proceedings from ISWC 2020*, pages 56–71.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Tsai, W.-H. S. and Chuan, C.-H. (2023). Humanizing chatbots for interactive marketing. *The Palgrave handbook of interactive marketing*, pages 255–273.