BRIEF REPORT

# REVISED Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful? [version 3; peer review: 3 approved]

Eduardo C. Garrido-Merchán[1], Jose Luis Arroyo-Barrigüete[1,2], Francisco Borrás-Pala[1], Leandro Escobar-Torres[1], Carlos Martínez de Ibarreta[1], Jose María Ortíz-Lozano[1,2], Antonio Rua-Vieites[1,2]

[1]Universidad Pontificia Comillas, Madrid, Community of Madrid, Spain
[2]Santalucía Chair of Analytics for Education, Madid, Spain, Spain

**v3** **First published:** 11 Jul 2024, **13**:791
https://doi.org/10.12688/f1000research.153129.1
**Second version:** 23 Sep 2024, **13**:791
https://doi.org/10.12688/f1000research.153129.2
**Latest published:** 17 Oct 2024, **13**:791
https://doi.org/10.12688/f1000research.153129.3

## Abstract

### Abstract

### Background

Large Language Models (LLMs), as in the case of OpenAI™ ChatGPT-4 ™ Turbo, are revolutionizing several industries, including higher education. In this context, LLMs can be personalised through customization process to meet the student demands on every particular subject, like statistics. Recently, OpenAI launched the possibility of customizing their model with a natural language web interface, enabling the creation of customised GPT versions deliberately conditioned to meet the demands of a specific task.

### Methods

This preliminary research aims to assess the potential of the customised GPTs. After developing a Business Statistics Virtual Professor (BSVP), tailored for students at the Universidad Pontificia Comillas, its behaviour was evaluated and compared with that of ChatGPT-4 Turbo. Firstly, each professor collected 15-30 genuine student questions from “Statistics and Probability” and “Business

**Open Peer Review**

**Approval Status** ✓ ✓ ✓

| | 1 | 2 | 3 |
|---|---|---|---|
| **version 3** (revision) 17 Oct 2024 | | | ✓ view |
| **version 2** (revision) 23 Sep 2024 | ✓ view | ✓ view | ? view |
| **version 1** 11 Jul 2024 | ? view | ? view | |

1. **Erik Carbajal-Degante**, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico
2. **María Beatriz Corchuelo Martínez-Azua**, Universidad de Extremadura,, Badajoz, Spain
3. **FX. Risang Baskara**, Universitas Sanata Dharma, Depox, Indonesia

Any reports and responses or comments on the article can be found at the end of the article.

Statistics" courses across seven degrees, primarily from second-year courses. These questions, often ambiguous and imprecise, were posed to ChatGPT-4 Turbo and BSVP, with their initial responses recorded without follow-ups. In the third stage, professors blindly evaluated the responses on a 0-10 scale, considering quality, depth, and personalization. Finally, a statistical comparison of the systems' performance was conducted.

**Results**

The results lead to several conclusions. Firstly, a substantial modification in the style of communication was observed. Following the instructions it was trained with, BSVP responded in a more relatable and friendly tone, even incorporating a few minor jokes. Secondly, when explicitly asked for something like, "I would like to practice a programming exercise similar to those in R practice 4," BSVP could provide a far superior response. Lastly, regarding overall performance, quality, depth, and alignment with the specific content of the course, no statistically significant differences were observed in the responses between BSVP and ChatGPT-4 Turbo.

**Conclusions**

It appears that customised assistants trained with prompts present advantages as virtual aids for students, yet they do not constitute a substantial improvement over ChatGPT-4 Turbo.

**Keywords**
Artificial Intelligence, ChatGPT, customisation, virtual instructor, higher education, statistics

**Corresponding author:** Jose Luis Arroyo-Barrigüete (jlarroyo@comillas.edu)

**Author roles: Garrido-Merchán EC**: Conceptualization, Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Arroyo-Barrigüete JL**: Conceptualization, Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Borrás-Pala F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Escobar-Torres L**: Data Curation, Formal Analysis, Writing – Review & Editing; **Martínez de Ibarreta C**: Data Curation, Formal Analysis, Writing – Review & Editing; **Ortíz-Lozano JM**: Data Curation, Formal Analysis, Writing – Review & Editing; **Rua-Vieites A**: Data Curation, Formal Analysis, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work is part of the NORIA research project (The Impact of Artificial Intelligence on the Legal Framework: Special Consideration of Its Effect on Legal Liabilityâ Grant number: PP2023_1, Universidad Pontificia Comillas). This work is also partially funded by the Santalucía Chair of Analytics for Education, at the Universidad Pontificia Comillas.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Garrido-Merchán EC, Arroyo-Barrigüete JL, Borrás-Pala F *et al.* **Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful? [version 3; peer review: 3 approved]** F1000Research 2024, **13**:791 https://doi.org/10.12688/f1000research.153129.3

**First published:** 11 Jul 2024, **13**:791 https://doi.org/10.12688/f1000research.153129.1

**REVISED** **Amendments from Version 2**

In this new version (version 3), we have carefully addressed the reviewer's insightful comments. It is worth noting that the feedback appears to reference version 1 of the manuscript, although we had already submitted version 2 by the time the review was conducted. This is likely due to the close timing between when we uploaded the new version and when the review was conducted. Many of the concerns raised, such as the need for a more comprehensive literature review, clearer explanations of the methodology, and a stronger discussion, had already been addressed in version 2. However, in version 3, we have further refined the manuscript to incorporate the remaining suggestions. Specifically, we have provided additional details on the evaluation criteria, enhanced the statistical analysis by including confidence intervals and applying a Bonferroni correction, and expanded our discussion of the practical implications for educators and institutions. We have also proposed new lines of future research to further strengthen the paper's contribution.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

The rapid advancements in statistical generative artificial intelligence (AI) (Murphy, 2023), particularly in the realm of natural language processing and generation with the emergence of Large Language Models (LLMs) (Gozalo-Brizuela and Garrido-Merchán, 2023b, Zhao et al., 2023), based on the transformers architecture, have given birth to a new paradigm in a plethora of sectors (Gozalo-Brizuela and Garrido-Merchán, 2023a), like marketing (Fraiwan and Khasawneh, 2023), higher education (Baskara, 2023; Sullivan et al., 2023) and research (Garrido-Merchán, 2023). Among the most notable developments in this field is OpenAI's ChatGPT-4 Turbo (OpenAI, 2023), a sophisticated language model that has demonstrated remarkable capabilities in generating human-like text (Garrido-Merchán et al., 2023) and performing several tasks accurately (Peng et al., 2023). This technology's potential in the educational sector, especially in creating virtual teaching assistants (Baidoo-Anu and Ansah, 2023), is immense. However, when customised for specific educational purposes, these AI models' effectiveness and practical utility remain burgeoning research areas.

Customised generative AI, particularly in LLMs like ChatGPT-4, involves configuring the model with specific data or prompts for tailored tasks, such as being a virtual instructor. This conditioning enhances its effectiveness in specialised roles, like serving as a virtual professor. OpenAI's new natural language interface for customization makes this process accessible across various fields. The relevance of this research stems from the growing demand for personalised learning in higher education. Customised AI models promise more engaging and personalised interactions, potentially transforming education. However, the true impact of these models on learning outcomes requires rigorous investigation to validate their effectiveness beyond marketing claims.

This study, therefore, focuses on evaluating the efficacy of a customised GPT version of ChatGPT-4 Turbo, developed as a Business Statistics Virtual Professor (BSVP), specifically for statistics students at the Business Faculty of Universidad Pontificia Comillas. By comparing the performance of this tailored model with the standard ChatGPT-4 Turbo in this particular task, this research aims to provide insights into the actual benefits and limitations of AI customisation in an educational context.

### Related work

The integration, challenges and opportunities of Generative AI into higher education, especially in the context of teaching, have garnered considerable attention in recent years (Michel-Villarreal et al., 2023). This section reviews the latest research in the field (Lo, 2023), emphasising studies that explore the role of generative AI in teaching, its application as a virtual assistant, and its contribution to academic research.

Recent studies in this domain have focused on the efficacy of generative AI in enhancing teaching methodologies (Baidoo-Anu and Ansah, 2023). These works highlight the potential of AI in personalising learning experiences, providing real-time feedback, and augmenting traditional teaching practices (Kasneci et al., 2023; Zhai, 2022). For example, ChatGPT has been proven helpful for lifelong learning (Rawas, 2023), as, for instance, it can readapt the teaching lessons to the latest advances of rapidly changing technologies.

However, generative AI has also raised a debate about evaluation methodologies of higher education (Anders, 2023), as students can use its content generation to cheat easily (Cotton et al., 2024). For example, evaluations done by professors have changed to adapt to this paradigm shift as, for instance, traditional assessments are easier to cheat than ever with generated content of Generative AI (Rudolph et al., 2023).

Another significant area of research involves using generative AI as virtual assistants in educational settings (Chheang et al., 2023). These studies explore the capabilities of AI assistants in managing student inquiries, offering personalised tutoring, and facilitating learning outside the traditional classroom environment (Ruiz-Rojas et al., 2023).

Finally, the role of generative AI in academic research (Xames and Shefa, 2023) has been an area of growing interest (Rahman and Watanobe, 2023). These investigations delve into how AI can assist in data analysis, brainstorming of ideas, literature review, synthetic data generation, text simplification and even in helping to write some sections of research papers, thereby augmenting the research capabilities of scholars and students alike (Garrido-Merchán, 2023).

## Generative Pretrained Transformers (GPTs)

The evolution of Generative Pretrained Transformers (GPTs) (Radford et al., 2018) has produced a paradigm shift in the democratisation of natural language processing (NLP) (Chowdhary and Chowdhary, 2020). The journey began with the original GPT model (Radford et al., 2018), introduced by OpenAI, whose novelty includes unsupervised learning to predict the next word in a sentence, not only supervised learning as was done before. More concretely, GPT's methodology encompassed a dual-phase process: an initial 'pre-training' stage using an unsupervised generative approach to establish baseline parameters through language modelling, followed by a customization stage, where these parameters were refined and tailored to a specific task in a supervised, discriminative manner.

This model laid the groundwork for more advanced iterations. GPT-2, developed by OpenAI (Radford et al., 2019) marked a significant leap with its 1.5 billion parameters and more engineering tricks, demonstrating enhanced text generation capabilities and enabling the hypothesis that scale was all that natural language processing needs. However, its behaviour showed clues of underfitting, being its capacity, despite its 1.5 billion parameters, which were too simple for the complexity of the Webtext corpus with 45 millions of webpages, as we illustrate in Figure 1.

We can clearly see how, according to this curve, the model is still in the underfitting regime. Recall that the overfitting regime starts when the test loss function curve surpasses the training set curve, indicating that the model is representing patterns that do not generalize outside of the sample which has been trained on. In order to reach the optimum point shown on Figure 2, a higher number of parameters was needed. This is the reason why GPT-3 and GPT-4 use a higher number of parameters, to try and reach this optimum point with respect to the WebText dataset, which is now more complex that in GPT-2 times and hence the model is going to require a higher capacity, because if it does not have the necessary capacity it will incur again in the underfitting issue that we have presented. We illustrate in Figure 2 an explanation of the underfitting zone suffered by the GPT-2 model and diagnosed by OpenAI researchers.

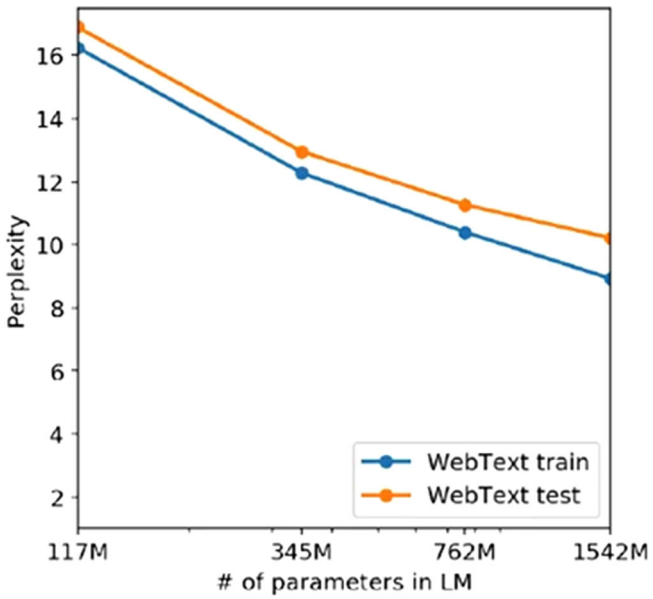


**Figure 1. Training and test set perplexities as a function of the millions of parameters of the GPT models (Source: Radford et al., 2019).**


**Figure 2. Underfitting and overfitting zones of a machine learning model with respect to its number of parameters given a dataset illustrated by the estimation of a particular loss function error in both datasets.** The underfiting issue appears when the model capacity is not able to represent the complexity of the data, incurring in a higher error than the one that can be obtained by increasing the number of parameters of the model, which was what happened with GPT models with respect to the WebText dataset (Source: own elaboration).

Motivated by this underfitting hypothesis, OpenAI launched GPT-3 (Brown et al., 2020), revolutionising the field with its 175 billion parameters and offering unprecedented language understanding and generation proficiency. It is important to emphasise that each iteration of GPT has built upon the transformer architecture (Vaswani et al., 2017). This architecture abandoned the recurrent layers used in previous models, relying instead on a self-attention mechanism that allowed the model to weigh the significance of different parts of the input data.

ChatGPT then emerged as a GPT 3.5 version that optimised the conversational experience with a user, being ChatGPT-4 (OpenAI, 2023) and ChatGPT-4 Turbo, standing out with its enhanced capabilities and efficiency, in comparison with GPT-3 (Peng et al., 2023). This version maintains the core transformer architecture but introduces several optimisations for speed and performance.

A critical component in developing GPT models, especially ChatGPT-4 Turbo that explains its outstanding behaviour is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). This training approach involves fine-tuning models based on feedback from human trainers. Initially, the model generates responses based on its pretraining; these responses are then evaluated by humans who provide ratings or improved versions of the responses. The model is subsequently retrained to prefer the human-approved responses. This method ensures that the model's outputs align more closely with human preferences, leading to more accurate and contextually appropriate responses that, now with the fine-tuned versions of ChatGPT like BSVP (Garrido-Merchán et al., 2024b), can gain even more importance.

The fine-tuning process in GPT models allows for the customisation of the base model to suit specific applications or domains. The fine-tuning process involves training the pre-existing model on a smaller, domain-specific dataset, enabling it to adapt its responses to the nuances of a particular field or user requirement. Fine-tuning can significantly enhance the model's performance in specialised tasks by adjusting its outputs to be more aligned with the specific content, style, or tone required by the application. This is precisely one of the advantages of its use in education.

The recent systematic review by Dong et al. (2024) highlights that research on the use of LLMs in education reveals both significant risks, such as obstruct the development of students' critical thinking skills or lead problems in academic integrity, as well as potential positive impacts on the learning process. Some authors argue that "Large language models, such as ChatGPT, have the potential to revolutionize teaching and assist in teaching processes. […]. [For example]

teachers can use large language models to create personalized learning experiences for their students" (Kasneci et al., 2023, p. 2). Specifically addressing their use as virtual assistants in higher education, several studies suggest that LLMs can support learning (Laato et al., 2023), contributing to personalized learning and knowledge access (Salem & Shaalan, 2024; Yigci et al., 2024). In fact, some research already indicates significant student use of these types of virtual assistants (Flores Limo et al., 2023). There are even proposals for chatbots specifically designed for higher education (Wang et al., 2023), which seem to perform better than ChatGPT on tasks related to course-specific content or less commonly known topics. This is precisely what we aim to assess, by comparing the standard version of ChatGPT with its customized version.

## Methods

Initially, a virtual assistant for Statistics courses taught at Universidad Pontificia Comillas was created. The assistant was instructed via prompt with specific directions regarding communication style. The decision was made to customize the model exclusively through prompt engineering, motivated by the intent to evaluate this new personalization feature offered by OpenAI. Prompt engineering allows for the adaptation of advanced language models such as ChatGPT without the need for complex technical interventions, thereby facilitating their use by individuals without specialized programming knowledge. This method promises to democratize the creation of personalized virtual assistants, making them accessible to a broader audience. Hence, there is a keen interest in assessing its effectiveness.

Additionally, contextual documentation was provided: two books written by three professors of the subject and signatories of this research (Borrás-Pala et al., 2019a, 2019b), as well as the R programming practices document, prepared by another three different professors, who are also authors of this work. Over three days, two authors tested the system, progressively refining the prompt until they achieved a version they considered acceptable.

The prompt utilized was designed to focus the model on key areas such as descriptive statistics, probability, and statistical inference. The prompt was structured with three priorities. The first priority was to ensure that responses were personalized, aligning with the way the subject matter is taught in the "Statistics and Probability" and "Business Statistics" courses at the Business Faculty of Universidad Pontificia Comillas. To achieve this, the model was instructed to always prioritize the content from the contextual documentation, with specific directives included. For instance, the instructions stated, "If asked about content related to descriptive statistics, probability, or inference, give absolute priority to the contents of the statistics books uploaded. Give them maximum weight and do not use other sources unless the prompt asks for content that is not contained in the books," and "If a student asks about a practice, consult the 'Programming Practices' document to respond. No other source is acceptable. Only that one." The second priority was the use of language appropriate for the average student at this university. Instructions incorporated for this purpose included "Use Spanish from Spain (Castilian)," "Do not digress, be concise," and "When you want to say 'assume,' use 'suppose.'" The third priority of the prompt was to employ a communication style that is engaging and relatable to the students. This was achieved by incorporating directives such as "Adopt the tone of an influencer who popularizes content. For example, be slightly enthusiastic in your responses, using emoticons in your explanations," and "At some point in your response, make a joke about the ICADE professor [...] to enhance the student's experience." These structured instructions were crafted to guide the model effectively, ensuring that its outputs were both academically aligned with the university's standards and engaging for the students, thereby optimizing the educational interaction.

Once the system was refined, the evaluation began. The study was conducted through the assessment of BSVP's response quality by the five professors who signed this work but did not participate in the generation and subsequent adjustment of the prompt. Specifically, the work was carried out in four different stages. Firstly, each professor collected between 15 and 30 questions posed by students of the 'Statistics and Probability' and 'Business Statistics' courses, which are taught across seven different degrees. A final sample of 136 questions was obtained. In most cases, these were second-year courses (mostly students aged 19-20) and, in some instances, third-year courses (mostly students aged 20-21). All questions had to be genuine inquiries made by students during classes or tutoring sessions. This is a highly relevant aspect, as students often struggle to clearly and precisely articulate their doubts (e.g., 'I don't understand what this Student's t is about'; 'In the Poisson binomial, how is lambda calculated?'): it's essential to evaluate the system's ability to respond to these kinds of questions competently, even if the formulation of the question itself is imprecise or even incorrect. If BSVP is to act as a virtual assistant for students, it should be able to answer such questions despite their ambiguity, lack of definition or even errors in the question itself. The questions collected are those that students typically ask in class or during tutoring sessions (not specifically for this study) and have been used anonymously. Intentionally, the questions collected by the professors were not coordinated, which implies that a few questions collected by one researcher might be similar to those collected by another. This occurred in some cases with questions that are very common among students. For example: 'I don't fully understand the difference between the intersection of two random events and one being conditioned on the other. 'or 'How can I tell if a problem is asking for the probability of an intersection or a conditioned event? 'In any case, since these were real questions, the wording was never identical,

allowing for the testing of both systems' (ChatGPT-4 Turbo and BSVP) ability to respond to different formulations. In the second stage, each question was posed to ChatGPT-4 Turbo and BSVP (Garrido-Merchán et al., 2024b), noting down both complete responses. To ensure comparability, there were no follow-up questions or clarifications; the first provided response was copied, whether satisfactory or not. In the third phase, the professors who had not participated in generating and adjusting the prompt evaluated the responses from ChatGPT-4 Turbo and BSVP, scoring them on a scale of 0 to 10. The choice of this specific scale responds to the characteristics of the Spanish university system, where it is the default scale used to evaluate university students. Therefore, the professors responsible for this evaluation are familiar with this scale. It is important to note that the evaluation was blind, as each professor assessed both responses without knowing who the author was (ChatGPT-4 Turbo or BSVP). Only the two professors who did not participate in the evaluation had this information. Specifically, three different dimensions were evaluated: quality of the response (clarity, conciseness, etc.); depth of the response (to what extent it is as complete as possible); and personalisation (degree of closeness to the way the subject is taught at the university where the study was conducted). To give a concrete example regarding the dimension of personalization, when teaching probability calculations for the normal distribution, most statistics textbooks rely on tables or statistical software. However, at our university, we emphasize that students develop the ability to perform approximate calculations without using tables or software, leveraging the properties of the normal distribution. That is, by knowing the percentage of data within the intervals of $\mu \pm \sigma$, $2\sigma$, and $3\sigma$, students should be able to estimate approximate probabilities. This is a distinction that BSVP should consider. Results are available at Garrido-Merchán et al. (2024a). Finally, in the fourth stage, a statistical comparison of the results obtained by both systems was carried out. Specifically, a paired samples t-test assuming equal variances was conducted for the mean differences in each of the three indicated dimensions.

## Results

Starting with a qualitative assessment, a substantial modification in the communication style was observed. As per its training, BSVP responded in a much more approachable and friendly tone. In fact, it often began responses with phrases like 'Dear ICADE student, …'[1] 'This question you ask is very interesting,' or 'Excellent question, my dear ICADE student!' The farewells were also more cordial ('a big hug,' 'I hope this has helped you'), and occasionally, they incorporated small jokes ('Perhaps your ICADE teacher might say something different, though I doubt it. But after all, they are human, and I am not, so I know much more than them')[2]. Greater conciseness in the responses was also generally observed, as instructed in the training prompt. A highly relevant aspect is that when explicitly asked for something like 'I would like to practice a programming exercise similar to those in R programming practice 3,' BSVP was capable of providing a much superior response: having access to contextual documentation, it was able to address the request, something that was not possible for ChatGPT-4 Turbo[3]. However, as a trade-off, the response times were generally longer. Regarding the content, a total of 136 questions were obtained, which, as mentioned, were evaluated according to three dimensions: quality, depth, and personalisation. Figure 3 shows the corresponding bar plots.

The comparative analysis of the performance of both systems (see Table 1) suggests no significant differences in any dimension: the p-values obtained across the three dimensions, all of which exceed 0.05, suggest that the observed differences may simply be due to random variations in the data sample rather than a systematic effect of the customization implemented in the BSVP. In fact, after applying the Bonferroni correction for multiple comparisons, the results become even clearer, with p-values of 0.82, 1, and 1 for Quality, Depth, and Personalisation, respectively. Additionally, to complement these results, the confidence intervals (CI) and the effect size (Cohen's d) has been calculated for each dimension (effsize). Once again, a negligible effect size is confirmed in all cases, significantly below the threshold of 0.2 typically used to denote small effects. These effect sizes reinforce the conclusion that the customization of the BSVP system has not resulted in significant improvements in student interaction compared to the standard ChatGPT-4 Turbo model. The most interesting aspect is the absence of differences in personalisation (the degree of closeness to the way the subject is taught at the university where the study was conducted), indicating that the contextual documentation has not served to offer adapted content. As mentioned, this documentation is handy when the question explicitly references course content (i.e., 'I would like to practice a programming exercise similar to those in R programming practice 3'), as it allows BSVP to respond competently. However, in more general questions like those included in this evaluation, which do not require the consultation of contextual documentation, there are no differences between BSVP and ChatGPT-4 Turbo.

[1]ICADE - Instituto Católico de Administración y Dirección de Empresas (Catholic Institute of Business Administration and Management). It is the name of the business school of the Universidad Pontificia Comillas, where the study was conducted.
[2]To ensure that the evaluation was blind, all these phrases were removed from the responses, so that the evaluators were not aware of them.
[3]Logically, questions of this nature were not included in the evaluation.

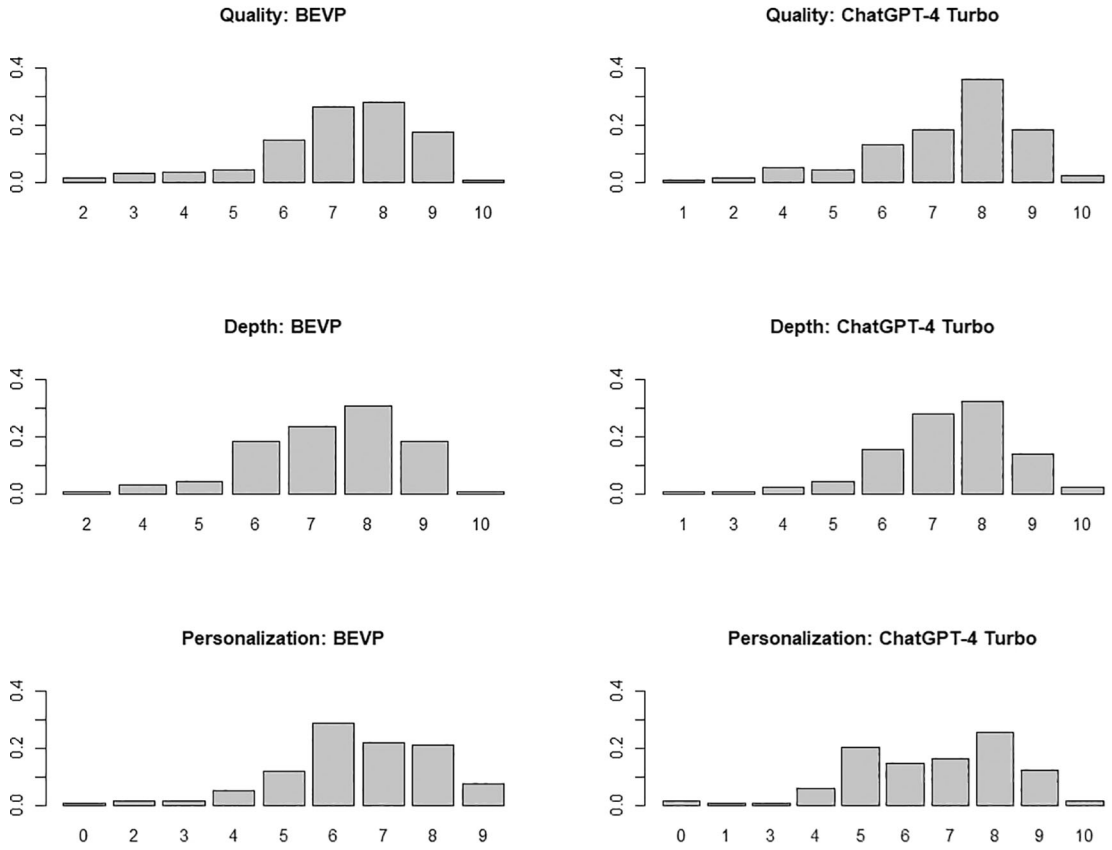


**Figure 3. Bar plot of the scores obtained by BSVP and ChatGPT-4 Turbo in each of the three dimensions analysed (figure generated with R).**

**Table 1. Results obtained in each dimension. Mean, standard deviation (sd), and t-test for mean difference.**

| | Quality | Depth | Personalisation |
|---|---|---|---|
| **BSVP: mean (sd)** | 7.12 (1.60) | 7.30 (1.36) | 6.50 (1.57) |
| **ChatGPT-4 Turbo: mean (sd)** | 7.30 (1.62) | 7.29 (1.39) | 6.64 (1.84) |
| **t-test** | t = -1.098<br>df = 135<br>p-value = 0.274<br>CI: [-0.51, 0.15]<br>effsize = -0.11 | t = 0.096<br>df = 135<br>p-value = 0.924<br>CI : [-0.29, 0.32]<br>effsize = 0.01 | t = -0.855<br>df = 135<br>p-value = 0.394<br>CI: [-0.46, 0.18]<br>effsize = -0.08 |

## Discussion

The main conclusions of this research can be summarised in three key ideas. Firstly, differences in communication style are indeed noticeable. Training via prompt has created a virtual assistant whose style is distinct from that of ChatGPT-4 Turbo. Secondly, BSVP has a significant advantage over ChatGPT-4 Turbo: its contextual documentation allows it to respond to specific course content queries, which ChatGPT-4 Turbo cannot do. This is not a minor aspect, as students often pose questions this way (e.g., 'Could you provide an example of a problem like those in chapter 4?'; 'I don't understand the first part of the R programming practice 6'). Lastly, regarding general content, no significant differences are evident. That is, ChatGPT-4 Turbo can answer any query like BSVP. However, we must consider that we are dealing with a subject that is quite basic and for which there is an enormous amount of information. Therefore, the responses cannot vary much in terms of quality and depth. Customisation via prompt seems to show specific improvements, especially if students prefer a friendlier communication style and targeted content queries. However, BSVP provides no benefit to students seeking doubt resolution over ChatGPT-4 Turbo.

On the other hand, as it is illustrated on the results section, the customized GPT version has shown a better performance in communication style with respect to the not customized version, which represents a critical advantage for users. The answer style provided by the standard GPT model may not be the usual way to communicate with students in different cultures, organizations and universities, depending on factors such as countries, different studies or beliefs. It is well known that university students typically interact between themselves in a specific manner (Gorsky et al., 2006), so this style can be introduced in the configuration prompt, making the model generate text in this fashion and not sound weird by students, which is a necessity for them (Jochim & Lenz-Kesekamp, 2024). By personalizing the model's communication style to align with that of the professor or the expectations of the organization or university, we can enhance the benefits that generative AI models provide to students (Tai & Chen, 2024). This personalization helps remove cultural barriers that students might face regarding the style of the texts generated by the model.

Regarding a potential improvement of performance by the BSVP customized GPT version with respect to the GPT-4 Turbo model, we do not empirically observe such behavior. Consequently, we hypothesize several causes that could be simultaneously affecting the behaviour of the BSVP customized GPT version. First, undergraduate business statistics is a subject with little dissent, in the sense that its syllabus is objective and very popular on the internet. Hence, our added specific theoretical materials of the subject do not add a significant amount of new knowledge to the GPT representation of information of its corpus encoded in its parameters. Observe that we are only describing here the theoretical content because, in the case of the practical content, if we do specific practices not done by the rest of the universities, then the customized GPT version can effectively provide unique answers as the result of its customization. We also hypothesize that if we had a subject with different schools of thought, such in the case of philosophy, for example, then, the performance of customized GPT models for education could be dramatic, as the customized GPT would be able to provide only the required answer for the subject that studies a particular school of thought. For example, if we are teaching a class about philosophy of mind, we could provide answers of both materialist or dualist beliefs by uploading files describing the schools a priori in the customization. Undoubtedly, regarding performance, the usefulness of customized GPT models in these cases would be superior than in the case of frequentist statistics.

Another critical advantage of the customized GPT versions with respect to the standard model is its usability and speed of use by the students. Instead of having to upload the subject materials to the model, students are provided with a customized version of the model that already contains the relevant subject materials. This version includes specific instructions on how to use the materials effectively, which have been carried out by the subject professor.

Hence, students are more likely to use this chatbot compared to one without the preloaded materials, as they will trust its content more, knowing that a professor has customized it. Additionally, the convenience of not needing to upload extra materials further increases its appeal. Recall that trusting generative AI is one of the issues of these systems that needs to be solved if chatbots are going to be widely used in education (Amoozadeh et al., 2024). Moreover, the student can use the chatbot to generate personalized problems similar to those in the subject, increasing trust in the tool. The student has greater confidence that these exercises are relevant for exam preparation, rather than being general problems that may not align with the course content. Furthermore, the generated exercises can vary in difficulty, effectively assisting students in mastering challenging concepts where existing exercises may be too complex to solve without further support. These exercises can be either analytical or coding-based, providing valuable help to non-STEM students, such as those in business programs, in overcoming the challenges of STEM subjects (Coe et al., 2008), such as statistics, particularly through practice in the R programming language.

In summary, our findings suggest that customizing AI assistants like BSVP may offer practical benefits for educators and educational institutions by enhancing student interaction through communication styles tailored to their cultural and academic expectations, saving instructors time by addressing frequent queries, and increasing student trust in a tool customized with course content. Moreover, in complex subjects or those with multiple theoretical approaches, customization can provide more precise and relevant responses, further enhancing its utility in educational settings.

The study's main limitation is its preliminary nature. To validate our findings, an experiment where students, as end-users of BSVP, assess both systems' responses is necessary. However, accurately assessing responses poses challenges; students probably would not be able to discriminate based on the veracity of the result: they might prefer brief answers over more accurate, complex ones; and could be influenced by the communication style, potentially skewing their judgements. Despite these obstacles, with well-designed experiments, we can further explore system differences from a student perspective and extend the research to more specialised, advanced subjects, which is what we propose as future lines of research. Another interesting research line would be to investigate whether, in subjects with different schools of thought (such as philosophy) the performance of customized GPT models is indeed substantially better than that of the general GPT, as we suspect might be the case.

## Ethical considerations

This work does not require approval from the ethics committee. The questions collected are those that students typically ask in class or during tutoring sessions (not specifically for this study) and have been used anonymously. Therefore, their consent is not required. Additionally, all professors who evaluated the quality of the responses are co-authors of this work and thus give their consent. In conclusion, approval from the ethics committee is not required. Ethical approval and participant consent were not applicable due to the nature of the study.

## Author contributions

Eduardo C. Garrido Merchán and Jose Luis Arroyo-Barrigüete contributed to the study conception and design. All authors performed material preparation, data collection, and analysis. The first draft of the manuscript was written by Eduardo C. Garrido Merchán and Jose Luis Arroyo-Barrigüete, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

## Data availability

Figshare: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?, https://doi.org/10.6084/m9.figshare.26039461.v1 (Arroyo-Barrigüete et al., 2024a).

The project contains the following underlying data:

- Data.xlsx. Rating on a scale from 0 to 10 of all responses evaluated according to the three considered dimensions (quality, depth, and personalization).

Figshare: Sample of provided responses: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?, 10.6084/m9.figshare.26965354.v1 (Arroyo-Barrigüete, 2024).

- This document includes a sample of responses supplied by BSVP and ChatGPT-4 Turbo. The Excel document indicates which of the responses (A or B) was provided by each of the two systems.



A preprint of the article can be found at https://arxiv.org/abs/2312.03728. Garrido-Merchán, E. C., Arroyo-Barrigüete, J. L., Borrás-Pala, F., Escobar-Torres, L., de Ibarreta, C. M., Ortiz-Lozano, J. M., & Rua-Vieites, A. (2023). Real Customization or Just Marketing: Are Customized Versions of Chat GPT Useful?. arXiv preprint arXiv:2312.03728.

## Extended data

Figshare: Questionary: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?, https://doi.org/10.6084/m9.figshare.26128669.v1 (Arroyo-Barrigüete et al., 2024b).



## References


Amoozadeh M, Daniels D, Nam D, *et al.*: **Trust in Generative AI among students: An exploratory study.** *In Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1.* 2024, March; pp. 67–73.

Anders BA: **Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking?** *Patterns.* 2023; **4**(3): 100692–100694.
Publisher Full Text

Arroyo-Barrigüete JL, Garrido-Merchán EC, Borrás-Pala F, *et al.*: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful? figshare. [Dataset]. 2024a.
Publisher Full Text

Arroyo-Barrigüete JL, Garrido-Merchán EC, Borrás-Pala F, *et al.*: Questionary: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful? figshare. [Dataset]. 2024b.
Publisher Full Text

Arroyo-Barrigüete JL: Sample of provided responses: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful? figshare. [Dataset]. 2024.
Publisher Full Text

Baidoo-Anu D, Ansah LO: **Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning.** *J. AI.* 2023; **7**(1): 52–62.
Publisher Full Text

Baskara R: **Exploring the implications of ChatGPT for language learning in higher education.** *Indonesian Journal of English Language Teaching and Applied Linguistics.* 2023; **7**(2): 343–358.

Borrás-Pala F, Martinez de Ibarreta C, Escobar-Torres L: *Estadística Empresarial en 101 ejemplos (volumen I).* EV Services; 2019a.

Borrás-Pala F, Martinez de Ibarreta C, Escobar-Torres L: *Estadística Empresarial en 101 ejemplos (volumen II).* EV Services; 2019b.

Brown T, Mann B, Ryder N, *et al.*: **Language models are few-shot learners.** *Adv. Neural Inf. Proces. Syst.* 2020; **33**: 1877–1901.

Chheang V, Marquez-Hernandez R, Patel M, *et al.*: **Towards anatomy education with generative AI-based virtual assistants in immersive virtual reality environments.** *arXiv preprint arXiv:2306.17278.* 2023.

Chowdhary K, Chowdhary KR: **Natural language processing.** *Fundamentals of artificial intelligence.* New Delhi: Springer; 2020;

pp. 603–649.
Publisher Full Text

Christiano PF, Leike J, Brown T, *et al.*: **Deep reinforcement learning from human preferences.** *Adv. Neural Inf. Proces. Syst.* 2017; **30**.

Coe R, Searle J, Barmby P, *et al.*: *Relative difficulty of examinations in different subjects.* Durham: CEM centre; 2008.

Cotton DR, Cotton PA, Shipway JR: **Chatting and cheating: Ensuring academic integrity in the era of ChatGPT.** *Innov. Educ. Teach. Int.* 2024; **61**(2): 228–239.
Publisher Full Text

Dong B, Bai J, Xu T, *et al.*: **Large Language Models in Education: A Systematic Review.** In 2024 *6th International Conference on Computer Science and Technologies in Education (CSTE).* IEEE; 2024; pp. 131–134.

Flores Limo FA, Hurtado Tiza DR, Mamani Roque M, *et al.*: **Personalized tutoring: ChatGPT as a virtual tutor for personalized learning experiences.** *Przestrzeń Społeczna (Social Space).* 2023; **23**(1): 293–312.

Fraiwan M, Khasawneh N: **A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions.** *arXiv preprint arXiv: 2305.00237.* 2023.

Garrido-Merchán EC, Arroyo-Barrigüete JL, Borrás-Pala F, *et al.*: **Survey data on "Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?".** *Figshare.* 2024a.
Publisher Full Text

Garrido-Merchán E, Arroyo-Barrigüete JL, Borrás-Pala F, *et al.*: **"Profesor Estadística Empresarial" (Version 1.0) [Software].** 2024b.
Reference Source

Garrido-Merchán: **Best uses of ChatGPT and Generative AI for computer science research.** *arXiv preprint arXiv:2311.11175.* 2023.

Garrido-Merchán EC, Arroyo-Barrigüete JL, Gozalo-Brihuela R: **Simulating HP Lovecraft horror literature with the ChatGPT large language model.** *arXiv preprint arXiv:2305.03429.* 2023.

Gorsky P, Caspi A, Trumper R: **Campus-based university students' use of dialogue.** *Stud. High. Educ.* 2006; **31**(1): 71–87.
Publisher Full Text

Gozalo-Brizuela R, Garrido-Merchán EC: **A survey of Generative AI Applications.** *arXiv preprint arXiv:2306.02781.* 2023a.

Gozalo-Brizuela R, Garrido-Merchan EC: **ChatGPT is not all you need. A State of the Art Review of large Generative AI models.** *arXiv preprint arXiv:2301.04655.* 2023b.

Jochim J, Lenz-Kesekamp VK: *Teaching and testing in the era of text-generative AI: exploring the needs of students and teachers.* Information and Learning Sciences; 2024.
Publisher Full Text

Kasneci E, Seßler K, Küchemann S, *et al.*: **ChatGPT for good? On opportunities and challenges of large language models for education.** *Learn. Individ. Differ.* 2023; **103**: 102274.
Publisher Full Text

Laato S, Morschheuser B, Hamari J, *et al.*: **AI-assisted learning with ChatGPT and large language models: Implications for higher education.** In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT).* IEEE; 2023; pp. 226–230.

Lo CK: **What is the impact of ChatGPT on education? A rapid review of the literature.** *Educ. Sci.* 2023; **13**(4): 410.
Publisher Full Text

Michel-Villarreal R, Vilalta-Perdomo E, Salinas-Navarro DE, *et al.*: **Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT.** *Educ. Sci.* 2023; **13**(9): 856.
Publisher Full Text

Murphy KP: *Probabilistic machine learning: Advanced topics.* MIT Press; 2023.

OpenAI: **GPT-4 technical report.** 2023.
Reference Source

Peng B, Li C, He P, *et al.*: **Instruction tuning with gpt-4.** *arXiv preprint arXiv: 2304.03277.* 2023.

Radford A, Narasimhan K, Salimans T, *et al.*: **Improving language understanding by generative pre-training.** *Preprint. Work in progress.* 2018.
Reference Source

Radford A, Wu J, Child R, *et al.*: **Language models are unsupervised multitask learners.** *OpenAI blog.* 2019; **1**(8): 9.

Rahman MM, Watanobe Y: **ChatGPT for education and research: Opportunities, threats, and strategies.** *Appl. Sci.* 2023; **13**(9): 5783.
Publisher Full Text

Rawas S: **ChatGPT: Empowering lifelong learning in the digital age of higher education.** *Educ. Inf. Technol.* 2023; 1–14.

Rudolph J, Tan S, Tan S: **ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?** *J. Appl. Learn. Teach.* 2023; **6**(1): 342–363.

Ruiz-Rojas LI, Acosta-Vargas P, De-Moreta-Llovet J, *et al.*: **Empowering Education with Generative Artificial Intelligence Tools: Approach with an Instructional Design Matrix.** *Sustainability.* 2023; **15**(15): 11524.
Publisher Full Text

Salem M, Shaalan K: **ChatGPT: Advancing Education with Virtual Assistants.** In: Hassanien AE, Zheng D, Zhao Z, Fan Z, editors. *Business Intelligence and Information Technology. BIIT 2023. Smart Innovation, Systems and Technologies.* Vol. 394. Singapore: Springer; 2024.
Publisher Full Text

Sullivan M, Kelly A, McLaughlan P: **ChatGPT in higher education: Considerations for academic integrity and student learning.** *J. Appl. Learn. Teach.* 2023; **6**(1): 1–10.
Publisher Full Text

Tai TY, Chen HHJ: **Improving elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions.** *Comput. Educ.* 2024; **220**: 105112.
Publisher Full Text

Vaswani A, Shazeer N, Parmar N, *et al.*: **Attention is all you need.** *Adv. Neural Inf. Proces. Syst.* 2017; **30**.

Wang K, Ramos J, Lawrence R: **ChatEd: a chatbot leveraging ChatGPT for an enhanced learning experience in higher education.** *arXiv preprint arXiv:2401.00052.* 2023.

Xames MD, Shefa J: **ChatGPT for research and publication: Opportunities and challenges.** *Available at SSRN 4381803.* 2023.

Yigci D, Eryilmaz M, Yetisen AK, *et al.*: **Ozcan A.: Large Language Model-Based Chatbots in Higher Education.** *Adv. Intell. Syst.* 2024: 2400429.
Publisher Full Text

Zhai X: **ChatGPT user experience: Implications for education.** *Available at SSRN 4312418.* 2022.

Zhao WX, Zhou K, Li J, *et al.*: **A survey of large language models.** *arXiv preprint arXiv:2303.18223.* 2023.

# Open Peer Review

## Current Peer Review Status: ✔ ✔ ✔

**Version 3**

Reviewer Report 21 October 2024

https://doi.org/10.5256/f1000research.173379.r332743



✔ **FX. Risang Baskara**
English Letters Department, Universitas Sanata Dharma, Depox, Indonesia

Dear Authors,
Thank you for the opportunity to review the revised version of your manuscript. I have carefully examined the changes made in response to previous reviewer comments and the overall quality of the paper. This report provides a comprehensive assessment of the revised manuscript.

1. Overview
The study presents a timely and relevant investigation into the effectiveness of customized versions of large language models (LLMs) in educational settings, specifically comparing a standard ChatGPT-4 Turbo model with a customized Business Statistics Virtual Professor (BSVP). The research question is well-defined, and the methodology is generally sound. The revisions have significantly improved the paper's clarity, depth, and scientific rigor.

2. Literature Review
The authors have successfully addressed the previous concern regarding the literature review. The inclusion of more recent and relevant studies on AI in education (e.g., Dong et al., 2024; Flores Limo et al., 2023; Laato et al., 2023) provides a stronger foundation for the research and better contextualizes the work within the rapidly evolving field.

3. Methodology
The methodology section has been substantially improved:
a) BSVP Customization: The detailed explanation of the prompt engineering process and the specific instructions used for customization greatly enhances the reproducibility of the study.
b) Evaluation Criteria: The clarification on the use of the 0-10 scale and how professors were instructed to assess responses is helpful.
c) Sample Size: The justification for the sample size of 136 questions is now adequately explained.

4. Statistical Analysis
The statistical analysis has been significantly enhanced:
a) Effect Sizes: The inclusion of Cohen's d for each dimension provides a clearer understanding of

the magnitude of differences.
b) Confidence Intervals: The addition of confidence intervals in Table 1 improves the interpretation of results.
c) Multiple Comparisons: The application of the Bonferroni correction for multiple comparisons strengthens the validity of the statistical analysis.

5. Results and Discussion
The expanded discussion section offers a more nuanced interpretation of the results:
a) Practical Implications: The authors have provided a thorough exploration of the practical implications for educators and educational institutions.
b) Limitations: The discussion of study limitations, particularly regarding the generalizability from a single subject area, is well-articulated.
c) Future Research: The proposed directions for future research, especially the suggestion to investigate customized GPT models in subjects with different schools of thought, are valuable.

6. Data Availability
The addition of sample responses from both BSVP and ChatGPT-4 Turbo enhances the reproducibility of the study.

Minor Points
- The explanation of the underfitting issue in GPT models is clear and well-illustrated.
- The rationale for not using BLEU and ROUGE metrics is well-argued and appreciated.

7. Conclusion
The revised manuscript represents a significant improvement over the previous version. The authors have been responsive to reviewer feedback and have made appropriate revisions to address all major concerns. The study now provides a more comprehensive and rigorous analysis of the potential and limitations of customized AI models in educational settings.

The revised paper also makes a valuable contribution to the field of AI in education. It is well-written, methodologically sound, and provides important insights into the effectiveness of customized LLMs in educational contexts. I believe it is now suitable for indexing and will be of interest to researchers and practitioners in the field of educational technology and AI.
Thank you for the opportunity to review this interesting and timely research.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Technology-Enhanced Language Learning (TELL)visibiliyBlended Learning Approaches in EFL ContextsvisibilitFlipped Classroom Methodologies for Language AcquisitionvisibilitArtificial Intelligence Applications in Language EducationvisibilitComputer-Assisted Language Learning (CALL)visibilitDigital Tools and Platforms for EFL TeachingvisibiliyInnovative Pedagogies in Second Language Acquisition

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

**Version 2**

Reviewer Report 04 October 2024

https://doi.org/10.5256/f1000research.172088.r326147



**Erik Carbajal-Degante**
Universidad Nacional Autonoma de Mexico, Mexico City, Mexico City, Mexico

I appreciate the authors' willingness to address the suggestions. I believe their work has significantly improved with the changes made. I give my approval for indexing of this article.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Artificial intelligence, computer vision, natural language processing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 01 October 2024

https://doi.org/10.5256/f1000research.172088.r326146



**María Beatriz Corchuelo Martínez-Azua**
Department of Economics, Universidad de Extremadura,, Badajoz, Extremadura, Spain

The authors have considered the recommendations made and the current version of the article has been significantly improved. It is considered suitable for indexing.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Economics, Teaching Innovation

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 26 September 2024

https://doi.org/10.5256/f1000research.172088.r326478



?

**FX. Risang Baskara**
English Letters Department, Universitas Sanata Dharma, Depox, Indonesia

Summary of the Article: This study investigates the effectiveness of customized versions of large language models (LLMs) in educational settings, specifically comparing a standard ChatGPT-4 Turbo model with a customized Business Statistics Virtual Professor (BSVP) designed for university students. The research aims to assess whether customized AI models offer significant advantages over general-purpose models in addressing student queries in a specific subject area. The study collected 136 genuine student questions from statistics courses, had both systems respond to these questions, and then had professors blindly evaluate the responses on quality, depth, and personalization. The results showed no statistically significant differences between the two systems, leading to the conclusion that while customized assistants may have some advantages, they do not constitute a substantial improvement over ChatGPT-4 Turbo for the tested scenario.
Detailed Responses to Review Questions:

1. Is the work clearly and accurately presented and does it cite the current literature? Answer: Partly

The paper is generally well-structured and presents the research clearly. However, there are areas for improvement:
a) Literature Review: While the paper cites relevant literature, the review could be more comprehensive, particularly regarding recent developments in AI for education. The authors should expand their literature review to include more recent studies on LLMs in educational contexts, especially those discussing customization efforts similar to their own.
b) Clarity of Presentation: Some sections, particularly the methodology and results, could benefit from more detailed explanations. For instance, the process of customizing the BSVP could be described more thoroughly.
c) Current Literature: The paper would benefit from including more up-to-date references on the use of AI in education, particularly studies published in the last 1-2 years, to better contextualize their work within the rapidly evolving field.
Recommendation: The authors should expand their literature review, incorporating more recent studies on AI in education. They should also provide more detailed explanations in the methodology and results sections.

1. Is the study design appropriate and is the work technically sound? Answer: Yes

The overall study design is appropriate for addressing the research question. The use of genuine student questions and blind evaluation by professors are strengths of the methodology. However, there are areas where the technical soundness could be improved:
a) Sample Size: The authors should justify the choice of 136 questions or acknowledge it as a potential limitation if it's considered small for robust statistical analysis.
b) Evaluation Criteria: While the use of a 0-10 scale for evaluation is explained, more detail on how professors were instructed to apply this scale would be beneficial.
c) Customization Process: More technical details on how BSVP was customized, including specific prompts or fine-tuning methods used, would enhance the study's reproducibility.
Recommendation: The authors should provide more details on their sample size justification,

evaluation criteria instructions, and the technical aspects of the BSVP customization process.

1. Are sufficient details of methods and analysis provided to allow replication by others?
   Answer: Partly

The paper provides a good overview of the methodology, but some crucial details are missing that would be necessary for full replication:
a) BSVP Customization: More specific information on how BSVP was customized, including the exact prompts or instructions used, would be essential for replication.
b) Evaluation Process: While the evaluation process is described, more details on how professors were instructed to assess the responses would be helpful.
c) Statistical Analysis: The statistical methods are described briefly, but more details on the specific tests used and any data preprocessing steps would aid replication.
Recommendation: The authors should provide a more detailed description of the BSVP customization process, including specific prompts used. They should also elaborate on the instructions given to professors for evaluation and provide more details on their statistical analysis methods.

1. If applicable, is the statistical analysis and its interpretation appropriate? Answer: Partly

The statistical analysis, while appropriate, is relatively basic and could be expanded:
a) Effect Size: In addition to p-values, the authors should consider reporting effect sizes to give a better sense of the magnitude of any differences.
b) Confidence Intervals: Including confidence intervals would provide more information about the precision of the estimates.
c) Multiple Comparisons: If multiple t-tests were performed, the authors should consider adjusting for multiple comparisons.
d) Interpretation: The interpretation of the statistical results could be more nuanced, discussing not just the lack of significant differences but also what the results suggest about the practical implications of customization.
Recommendation: The authors should expand their statistical analysis to include effect sizes and confidence intervals. They should also consider adjusting for multiple comparisons if applicable and provide a more nuanced interpretation of their results.

1. Are all the source data underlying the results available to ensure full reproducibility?
   Answer: Yes

The authors have made their data available on Figshare, which is commendable for ensuring reproducibility. However, to further enhance reproducibility:
a) Data Description: A more detailed description of the dataset, including variable definitions and any data cleaning steps, would be helpful.
b) Code Availability: If any custom code was used for analysis, making this available would further aid reproducibility.
Recommendation: While the data is available, the authors should provide a more detailed description of the dataset and consider making any custom analysis code available.

1. Are the conclusions drawn adequately supported by the results? Answer: Partly

The main conclusions are supported by the results, but there are areas where the discussion and conclusions could be strengthened:
a) Implications: The authors could provide a more in-depth discussion of the implications of their findings for the field of AI in education.
b) Limitations: A more thorough discussion of the study's limitations, including the generalizability of results from a single subject area, would strengthen the paper.
c) Future Research: More specific suggestions for future research directions based on their

findings would be valuable.
d) Practical Applications: The authors could expand on the practical implications of their findings for educators and educational institutions considering the use of customized AI assistants.
Recommendation: The authors should expand their discussion section to more thoroughly explore the implications and limitations of their findings. They should also provide more specific recommendations for future research and discuss practical applications of their results.
Conclusion: This paper addresses an important and timely topic in the field of AI and education. While it provides valuable insights, there are several areas where the methodology, analysis, and discussion could be strengthened. The most critical points that must be addressed to make the article scientifically sound are:

1. Expanding the literature review to include more recent, relevant studies.
2. Providing more detailed information on the BSVP customization process.
3. Enhancing the statistical analysis with effect sizes and confidence intervals.
4. Offering a more nuanced interpretation of the results and their implications.

By addressing these points, along with the other suggestions provided, the authors can significantly improve the scientific rigor and impact of their study.

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Technology-Enhanced Language Learning (TELL)visibiliyBlended Learning Approaches in EFL ContextsvisibilitFlipped Classroom Methodologies for Language AcquisitionvisibilitArtificial Intelligence Applications in Language EducationvisibilitComputer-Assisted Language Learning (CALL)visibilitDigital Tools and Platforms for EFL TeachingvisibiliyInnovative Pedagogies in Second Language Acquisition

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have**

**significant reservations, as outlined above.**

Author Response 11 Oct 2024

**Jose Luis Arroyo-Barrigüete**

Detailed Responses to Review Questions:
**Is the work clearly and accurately presented and does it cite the current literature?**
Answer: Partly
The paper is generally well-structured and presents the research clearly. However, there are areas for improvement:
a) Literature Review: While the paper cites relevant literature, the review could be more comprehensive, particularly regarding recent developments in AI for education. The authors should expand their literature review to include more recent studies on LLMs in educational contexts, especially those discussing customization efforts similar to their own.
b) Clarity of Presentation: Some sections, particularly the methodology and results, could benefit from more detailed explanations. For instance, the process of customizing the BSVP could be described more thoroughly.
c) Current Literature: The paper would benefit from including more up-to-date references on the use of AI in education, particularly studies published in the last 1-2 years, to better contextualize their work within the rapidly evolving field.
**Recommendation:** The authors should expand their literature review, incorporating more recent studies on AI in education. They should also provide more detailed explanations in the methodology and results sections.
**Response:** We appreciate the reviewer's meticulous evaluation of our paper. We have noticed that, although the journal's website indicates that this review pertains to version 2 of the manuscript, it actually refers to the initial version (version 1). This is likely due to the close timing between when we uploaded the new version and when the review was conducted.
The new version (version 2) incorporated the comments from the first two reviewers. Consequently, many of the issues mentioned (which were indeed weaknesses of the initial version) have already been addressed. For example, the need for a more extensive and recent literature review (points a and c) and a better explanation of the methodological and results sections (point b) have been addressed. Both aspects were also mentioned by the other two reviewers and have been incorporated into the version 2. You can find a detailed explanation of these improvements in our responses to the first two reviewers.
However, as described below, some of the reviewer's comments had not been previously suggested. Therefore, we have incorporated them into the new version (version 3) that we are submitting along with this response letter.

**Recommendation: Is the study design appropriate and is the work technically sound?**
Answer: Yes
The overall study design is appropriate for addressing the research question. The use of genuine student questions and blind evaluation by professors are strengths of the methodology. However, there are areas where the technical soundness could be improved:
a) Sample Size: The authors should justify the choice of 136 questions or acknowledge it as a potential limitation if it's considered small for robust statistical analysis.
**Response:** We conducted an initial sample size estimation aiming for a power of 0.8, a p-

value of 0.05, and to detect an effect size of 0.25 in a two-tailed paired t-test. This resulted in a required sample of 128 questions. However, we decided to evaluate 136 questions to account for the possibility of encountering issues that might lead to discarding some data. Ultimately, all questions were valid, so our final sample consisted of 136 questions.

**Recommendation:** b) Evaluation Criteria: While the use of a 0-10 scale for evaluation is explained, more detail on how professors were instructed to apply this scale would be beneficial.
**Response:** The reason for this is the grading system used in Spanish universities. In this system, both assignments and exams are evaluated on a scale from 0 to 10: scores from 0 to 4.99 are failing grades, 5 to 6.99 are passing grades, 7 to 8.99 are considered "good," and 9 to 10 are "excellent." Consequently, university professors are comfortable using this scoring scale because it is the scale they use when evaluating students. We included this explanation in version 2: “The choice of this specific scale responds to the characteristics of the Spanish university system, where it is the default scale used to evaluate university students. Therefore, the professors responsible for this evaluation are familiar with this scale.”

**Recommendation:** c) Customization Process: More technical details on how BSVP was customized, including specific prompts or fine-tuning methods used, would enhance the study's reproducibility.
**Response:** This comment was also raised previously, and we addressed it in version 2 by adding the following explanation:
“*The decision was made to customize the model exclusively through prompt engineering, motivated by the intent to evaluate this new personalization feature offered by OpenAI. Prompt engineering allows for the adaptation of advanced language models such as ChatGPT without the need for complex technical interventions, thereby facilitating their use by individuals without specialized programming knowledge. This method promises to democratize the creation of personalized virtual assistants, making them accessible to a broader audience. Hence, there is a keen interest in assessing its effectiveness.*
*[...]*
*The prompt utilized was designed to focus the model on key areas such as descriptive statistics, probability, and statistical inference. The prompt was structured with three priorities. The first priority was to ensure that responses were personalized, aligning with the way the subject matter is taught in the "Statistics and Probability" and "Business Statistics" courses at the Business Faculty of Universidad Pontificia Comillas. To achieve this, the model was instructed to always prioritize the content from the contextual documentation, with specific directives included. For instance, the instructions stated, "If asked about content related to descriptive statistics, probability, or inference, give absolute priority to the contents of the statistics books uploaded. Give them maximum weight and do not use other sources unless the prompt asks for content that is not contained in the books," and "If a student asks about a practice, consult the 'Programming Practices' document to respond. No other source is acceptable. Only that one." The second priority was the use of language appropriate for the average student at this university. Instructions incorporated for this purpose included "Use Spanish from Spain (Castilian)," "Do not digress, be concise," and "When you want to say 'assume,' use 'suppose.'" The third priority of the prompt was to employ a communication style that is engaging and relatable to the students. This was achieved by incorporating directives such as "Adopt the tone of an influencer who*

*popularizes content. For example, be slightly enthusiastic in your responses, using emoticons in your explanations," and "At some point in your response, make a joke about the ICADE professor [...] to enhance the student's experience." These structured instructions were crafted to guide the model effectively, ensuring that its outputs were both academically aligned with the university's standards and engaging for the students, thereby optimizing the educational interaction."*

**Recommendation: Are sufficient details of methods and analysis provided to allow replication by others?**
Answer: Partly
The paper provides a good overview of the methodology, but some crucial details are missing that would be necessary for full replication:
a) BSVP Customization: More specific information on how BSVP was customized, including the exact prompts or instructions used, would be essential for replication.
**Response:** As noted earlier, we included a more detailed explanation on this matter in version 2.

**Recommendation:** b) Evaluation Process: While the evaluation process is described, more details on how professors were instructed to assess the responses would be helpful.
**Response:** The instructors were directed to assess the responses as follows. For evaluating the dimensions of quality (such as clarity and conciseness) and depth (the extent to which the response is as complete as possible), they were instructed to use the same criteria they would apply when grading student work. Regarding the dimension of personalization (the degree to which the response aligns with how the subject is taught at the university where the study was conducted), they were to assess whether the response was consistent with the specific content and teaching methods used at Universidad Pontificia Comillas.
To illustrate with a concrete example, when teaching probability calculations for the normal distribution, most statistics textbooks rely on tables or statistical software. However, at our university, we emphasize that students develop the ability to perform approximate calculations without using tables or software, leveraging the properties of the normal distribution. That is, by knowing the percentage of data within the intervals of $\mu \pm \sigma$, $2\sigma$, and $3\sigma$, students should be able to estimate approximate probabilities. This is a distinction that BSVP should consider; when addressing a question related to the normal distribution, it should provide a response aligned with this method of calculation. The same applies to other topics that have slight differences compared to how they are presented in other statistics textbooks. We have added a clarification on this matter in version 3.

**Recommendation:** c) Statistical Analysis: The statistical methods are described briefly, but more details on the specific tests used and any data preprocessing steps would aid replication.
**Response:** Since the evaluation was conducted by the authors themselves, data preprocessing was unnecessary; the scores assigned to the responses were error-free. However, regarding the methodology used, we have included an additional clarification to the paper. We specified that the paired samples t-tests were performed assuming equal variances because, in all three cases, Levene's test confirmed that there were no statistically significant differences between the variances of the two groups (p-value > 0.05).

**Recommendation: If applicable, is the statistical analysis and its interpretation**

**appropriate?**
Answer: Partly
The statistical analysis, while appropriate, is relatively basic and could be expanded:
a) Effect Size: In addition to p-values, the authors should consider reporting effect sizes to give a better sense of the magnitude of any differences.
**Response:** We agree that including effect sizes is important. We have already incorporated this into version 2 (see Table 1).

**Recommendation:** b) Confidence Intervals: Including confidence intervals would provide more information about the precision of the estimates.
**Response:** We have included the confidence intervals in Table 1.

**Recommendation:** c) Multiple Comparisons: If multiple t-tests were performed, the authors should consider adjusting for multiple comparisons.
**Response:** We completely agree with this point. We did not include it in the previous version because, since no significant differences were found, we considered the correction unnecessary. However, we have now incorporated it into version 3 by applying a Bonferroni correction to the p-values obtained.

**Recommendation:** d) Interpretation: The interpretation of the statistical results could be more nuanced, discussing not just the lack of significant differences but also what the results suggest about the practical implications of customization.
**Response:** We acknowledge that this was a weakness in the initial version. However, in version 2, the reviewer can observe that the discussion section has been substantially expanded.

**Recommendation: Are all the source data underlying the results available to ensure full reproducibility?**
Answer: Yes
The authors have made their data available on Figshare, which is commendable for ensuring reproducibility. However, to further enhance reproducibility:
a) Data Description: A more detailed description of the dataset, including variable definitions and any data cleaning steps, would be helpful.
**Response:** In version 2, we have already included additional information about the data used in the Figshare repository, as requested by one of the initial reviewers.

**Recommendation:** b) Code Availability: If any custom code was used for analysis, making this available would further aid reproducibility.
**Response:** The statistical analysis is actually quite simple and can be performed using any statistical software. We used R simply because it is the standard tool in our research, not because complex calculations were necessary. The computations are indeed straightforward. For example, the command to calculate a t-test is as follows:
t.test(PEE_Personalization, GPT_Personalization, var.equal = TRUE, paired = TRUE)

**Recommendation: Are the conclusions drawn adequately supported by the results?**
Answer: Partly
The main conclusions are supported by the results, but there are areas where the

discussion and conclusions could be strengthened:
a) Implications: The authors could provide a more in-depth discussion of the implications of their findings for the field of AI in education.
**Response:** As previously mentioned, in version 2 the discussion section was substantially expanded, which includes a more in-depth discussion of the implications of our findings.

**Recommendation:** b) Limitations: A more thorough discussion of the study's limitations, including the generalizability of results from a single subject area, would strengthen the paper.
**Response:** We completely agree with this comment. For that reason, in version 2 we specifically addressed this point. Our opinion is that if we had a subject with different schools of thought, such in the case of philosophy, for example, then, the performance of customized GPT models for education could be dramatic, as the customized GPT would be able to provide only the required answer for the subject that studies a particular school of thought. For example, if we are teaching a class about philosophy of mind, we could provide answers of both materialist or dualist beliefs by uploading files describing the schools a priori in the customization. Undoubtedly, regarding performance, the usefulness of customized GPT models in these cases would be superior than in the case of frequentist statistics.

**Recommendation:** c) Future Research: More specific suggestions for future research directions based on their findings would be valuable.
**Response:** We have incorporated a new future line of research related to the previous point: “Another interesting research line would be to investigate whether, in subjects with different schools of thought (such as philosophy) the performance of customized GPT models is indeed substantially better than that of the general GPT, as we suspect might be the case.”

**Recommendation:** d) Practical Applications: The authors could expand on the practical implications of their findings for educators and educational institutions considering the use of customized AI assistants.
**Response:** We have included an additional paragraph on this matter: “In summary, our findings suggest that customizing AI assistants like BSVP may offer practical benefits for educators and educational institutions by enhancing student interaction through communication styles tailored to their cultural and academic expectations, saving instructors time by addressing frequent queries, and increasing student trust in a tool customized with course content. Moreover, in complex subjects or those with multiple theoretical approaches, customization can provide more precise and relevant responses, further enhancing its utility in educational settings.”

**Recommendation:** Conclusion: This paper addresses an important and timely topic in the field of AI and education. While it provides valuable insights, there are several areas where the methodology, analysis, and discussion could be strengthened. The most critical points that must be addressed to make the article scientifically sound are:
Expanding the literature review to include more recent, relevant studies.
Providing more detailed information on the BSVP customization process.
Enhancing the statistical analysis with effect sizes and confidence intervals.

Offering a more nuanced interpretation of the results and their implications.
By addressing these points, along with the other suggestions provided, the authors can significantly improve the scientific rigor and impact of their study.
**Response:** We appreciate the reviewer's insightful comments and suggestions. We believe that the new version of our manuscript has addressed the weaknesses highlighted

***Competing Interests:*** No competing interests were disclosed.

**Version 1**

Reviewer Report 04 September 2024

https://doi.org/10.5256/f1000research.167979.r317573



?

**María Beatriz Corchuelo Martínez-Azua**
Department of Economics, Universidad de Extremadura,, Badajoz, Extremadura, Spain

Thank you for the opportunity to review the paper "Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?". This research examines the effectiveness of generative language models, focusing on ChatGPT, in education. It compares ChatGPT with a customized assistant, the Business Statistics Virtual Professor (BSVP), designed for students at Universidad Pontificia Comillas. The study assesses the potential of these AI models to enhance learning, highlighting the need for effective personalization and fine-tuning. The findings reveal no significant differences between the two methods, suggesting that personalization may not be as impactful as anticipated. The study concludes that while personalized AI models hold promise, further research is needed to optimize their use in education.
The following are some comments and suggestions that could enhance the overall impact of the research conducted:
The work is presented clearly and precisely, with the study's methodology, results, and implications detailed effectively. The conclusions are well-supported by the results, and the study references relevant works, including previous studies and the context of the Spanish university system, which strengthens the evaluation criteria used in the research. Howev**er,** the provided excerpts lack specific references to recent literature beyond the authors' previous work. To broaden the study's perspective, it would be beneficial to include references that discuss the use of language models in diverse educational contexts, both locally and internationally. This could offer a more comprehensive view of the applicability of the findings.
The study commendably provides access to the source data underlying the results, with data availability on Figshare and appropriate licensing that supports transparency and reproducibility. However, the research's impact could be further enhanced by offering more detailed descriptions of the dataset, which would improve its usability for future research.

While statistical results are presented, the study would benefit from a more in-depth discussion of their significance and practical implications. For instance, exploring how the results might affect educational practice or influence the adoption and use of virtual assistants in higher education would be valuable. Such an analysis could provide readers with a clearer understanding of the real-world applications and potential impact of the findings.

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Economics, Teaching Innovation

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 17 Sep 2024

**Jose Luis Arroyo-Barrigüete**

We sincerely thank the reviewer for the time dedicated to evaluating our research. Below, we provide responses to her comments.

1. **The work is presented clearly and precisely, with the study's methodology, results, and implications detailed effectively. The conclusions are well-supported by the results, and the study references relevant works, including previous studies and the context of the Spanish university system, which strengthens the evaluation criteria used in the research. However, the provided excerpts lack specific references to recent literature beyond the authors' previous work. To broaden the study's perspective, it would be beneficial to include references that discuss the use of language models in diverse educational contexts, both locally and internationally. This could offer a more**

**comprehensive view of the applicability of the findings.**

We sincerely appreciate your comment. Due to the space limitations required by the journal for a Brief Report, it is not possible to include a large number of new references or text. However, we have reviewed a significant number of papers on the subject and selected those that we believe provide relevant information for the present research. These have been included in the paper at the end of the "Related Work" section:

- Dong, B., Bai, J., Xu, T., & Zhou, Y. (2024, April). Large Language Models in Education: A Systematic Review. In 2024 6th International Conference on Computer Science and Technologies in Education (CSTE) (pp. 131-134). IEEE.
- Flores Limo, F. A.,... & Arias Gonzáles, J. L. (2023). Personalized tutoring: ChatGPT as a virtual tutor for personalized learning experiences. Przestrzeń Społeczna (Social Space), 23(1), 293-312.
- Laato, S., Morschheuser, B., Hamari, J., & Björne, J. (2023, July). AI-assisted learning with ChatGPT and large language models: Implications for higher education. In 2023 IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 226-230). IEEE.
- Salem, M., Shaalan, K. (2024). ChatGPT: Advancing Education with Virtual Assistants. In: Hassanien, A.E., Zheng, D., Zhao, Z., Fan, Z. (eds) Business Intelligence and Information Technology. BIIT 2023. Smart Innovation, Systems and Technologies, vol 394. Springer, Singapore. https://doi.org/10.1007/978-981-97-3980-6_25
- Wang, K., Ramos, J., & Lawrence, R. (2023). ChatEd: a chatbot leveraging ChatGPT for an enhanced learning experience in higher education. arXiv preprint arXiv:2401.00052.
- Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2024). Large Language Model-Based Chatbots in Higher Education. Advanced Intelligent Systems, 2400429. https://doi.org/10.1002/aisy.202400429

**2. The study commendably provides access to the source data underlying the results, with data availability on Figshare and appropriate licensing that supports transparency and reproducibility. However, the research's impact could be further enhanced by offering more detailed descriptions of the dataset, which would improve its usability for future research.**

In order to offer more detailed descriptions of the dataset, we have uploaded a sample of the responses provided by both BSVP and ChatGPT-4 Turbo to a public repository. We have included the link to these files in the data availability statement:

Figshare: Sample of provided responses: Real Customization or Just Marketing: Are Customized Versions of Generative AI Useful?,
https://doi.org/10.6084/m9.figshare.26965354.v1

This document includes a sample of responses supplied by BSVP and ChatGPT-4 Turbo.

**While statistical results are presented, the study would benefit from a more in-depth discussion of their significance and practical implications. For instance, exploring how the results might affect educational practice or influence the adoption and use of virtual assistants in higher education would be valuable. Such an analysis could provide readers with a clearer understanding of the real-world applications and potential impact of the findings.**

After checking our document we definitely agree with this observation and feel very grateful with the reviewer to have told us to provide a wider discussion section. Please observe that we have augmented the discussion section regarding how the customized model can impact the students use of generative AI tools. In particular, we have focused our analysis in how does communication style conditions the student use of these models and how it also influences the fact that the customization has been done by a professor of the subject, clearly modifying the trust that the students feel with respect to the generative AI system once that they know that it is, somehow, validated and explored by the professors of the subject. Moreover, we have also given an argument on why, regarding statistics, the performance of the customized version does not clearly outperform the standard version, based on the fact that undergraduate statistics is a popular topic on the internet and it is knowledge accepted by all communities. We hypothesize that for another more subjective topics with different schools of thought, like for example different fields of philosophy like philosophy of mind, the personalized answers can outperform the standard answers as they will focus only on the school of thought being taught in the subject. Finally, in this discussion, we also provided some additional examples and references to support our arguments.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 13 August 2024

https://doi.org/10.5256/f1000research.167979.r302467



?

**Erik Carbajal-Degante**
Universidad Nacional Autonoma de Mexico, Mexico City, Mexico City, Mexico

**Summary**:
The present study conducts a comparative analysis to identify significant differences between the widely recognized Large Language Model ChatGPT and a customized GPT-based assistant, specifically designed for tailoring question-and-answer tasks for students within a particular subject at a specific university. The clarity and organization of the paper structure is good. Please, find bellow my comments:

**Comments:**

- In the section on GPTs, I recommend that the authors clarify the concept of underfitting and provide a detailed explanation of the specific indicators or evidence of underfitting they are referring to in relation to the GPTs family.
- The process of fine-tuning involves a comprehensive adjustment of the model's architecture, hyperparameters, and learning mechanisms to ensure effective assimilation of the new information during inference. In light of this, authors should expand on this topic

to clarify whether a specific fine-tuning method was applied in the development of the BSVP assistant, or if the customization was limited to prompt engineering.
- In the Methods section, I recommend providing evidence of the types of instructions given via prompts for the BSVP assistant.
- In the Results section, this study could be enhanced by incorporating additional NLP metrics, such as BLEU and ROUGE, which provide reliable and complementary data alongside the expert evaluations from the professors. Including these metrics would significantly increase the paper's value by adding a quantitative dimension to the analysis.
- The statistical results indicate no significant difference between the outcomes produced by the GPT and BSVP methods. I recommend that the authors provide a more in-depth analysis of the interpretation of their results, particularly regarding the insights provided by the p-value.

In my opinion, addressing these comments would significantly improve the quality of this research, enhancing its suitability for indexing.

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Artificial intelligence, computer vision, natural language processing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 17 Sep 2024

**Jose Luis Arroyo-Barrigüete**

First of all, we would like to thank the reviewer for their comments and the time dedicated to evaluating our work. We will now proceed to respond to the issues raised.

**In the section on GPTs, I recommend that the authors clarify the concept of underfitting and provide a detailed explanation of the specific indicators or evidence of underfitting they are referring to in relation to the GPTs family**
We thank the reviewer for coming up with such an interesting topic. The underfitting issue was diagnosed by OpenAI scientists as the loss function of the training and test sets of GPT-2 with respect to the WebText corpus did not converge, as we illustrate in the paper in the section on GPTs. We also provided a very detailed explanation on the underfitting issue with a Figure where it can clearly be seen how the loss functions of GPT-2 were on the underfitting zone.

**The process of fine-tuning involves a comprehensive adjustment of the model's architecture, hyperparameters, and learning mechanisms to ensure effective assimilation of the new information during inference. In light of this, authors should expand on this topic to clarify whether a specific fine-tuning method was applied in the development of the BSVP assistant, or if the customization was limited to prompt engineering.**
We have explained in more detail this issue, adding the following text:
*The decision was made to customize the model exclusively through prompt engineering, motivated by the intent to evaluate this new personalization feature offered by OpenAI. Prompt engineering allows for the adaptation of advanced language models such as ChatGPT without the need for complex technical interventions, thereby facilitating their use by individuals without specialized programming knowledge. This method promises to democratize the creation of personalized virtual assistants, making them accessible to a broader audience. Hence, there is a keen interest in assessing its effectiveness.*

**In the Methods section, I recommend providing evidence of the types of instructions given via prompts for the BSVP assistant.**
We appreciate the comment and agree that it is important to expand on this point. Therefore, we have added the following text:
*The prompt utilized was designed to focus the model on key areas such as descriptive statistics, probability, and statistical inference. The prompt was structured with three priorities. The first priority was to ensure that responses were personalized, aligning with the way the subject matter is taught in the "Statistics and Probability" and "Business Statistics" courses at the Business Faculty of Universidad Pontificia Comillas. To achieve this, the model was instructed to always prioritize the content from the contextual documentation, with specific directives included. For instance, the instructions stated, "If asked about content related to descriptive statistics, probability, or inference, give absolute priority to the contents of the statistics books uploaded. Give them maximum weight and do not use other sources unless the prompt asks for content that is not contained in the books," and "If a student asks about a practice, consult the 'Programming Practices' document to respond. No other source is acceptable. Only that one." The second priority was the use of language appropriate for the average student at this university. Instructions incorporated for this purpose included "Use Spanish from Spain (Castilian)," "Do not digress, be concise," and "When you want to say 'assume,' use 'suppose.'" The third priority of the prompt was to employ a communication style that is engaging and relatable to the students. This was achieved by incorporating directives such as "Adopt the tone of an influencer who popularizes content. For example, be slightly enthusiastic in your responses, using emoticons in*

*your explanations," and "At some point in your response, make a joke about the ICADE professor [...] to enhance the student's experience." These structured instructions were crafted to guide the model effectively, ensuring that its outputs were both academically aligned with the university's standards and engaging for the students, thereby optimizing the educational interaction.*

**In the Results section, this study could be enhanced by incorporating additional NLP metrics, such as BLEU and ROUGE, which provide reliable and complementary data alongside the expert evaluations from the professors. Including these metrics would significantly increase the paper's value by adding a quantitative dimension to the analysis.**

We a priori agree with the reviewer in that providing a quantitative measure of the quality of the generated text is such an interesting idea. In fact, it is an idea that we will research in the short future. The problem is that current measures as BLEU and ROUGE only determine the syntactic similarity of the text being generated with a reference text. We provide a full explanation to the reviewer about why using BLEU and ROUGE is not the best idea in this particular case scenario. However, we feel very grateful because this comment has inspired us an exciting new research line to design better quantitative measure in the era of generative AI. Now, we provide the explanation:

In the domain of natural language processing, traditionally, BLEU and ROUGE have been widely used metrics to assess the quality of machine-generated text by comparing it with reference texts. However, the application of these metrics in evaluating AI-generated content for technical or academic fields, such as the one covered in this paper, presents significant limitations that we illustrate in this subsection, making them a biased and not representative measure of evaluating the quality of the texts generated by AI, than for this particular scenario can only be evaluated by a expert commitee. In order to develop our argument we present an illustrative example. Consider the following reference text:
"Poisson Regression models are best used for modeling events where the outcomes are counts. Or, more specifically, count data: discrete data with non-negative integer values that count something, like the number of times an event occurs during a given timeframe or the number of people in line at the grocery store."
Now, to illustrate why quantitative measures such as BLEU and ROUGE are biased and do not represent the real loss function of this problem. We generate two texts with the generative AI. One of those texts is almost identical to the reference text, but the concept is wrong, as it considers that the Poisson distribution must be used for continuous variables. The other text is significantly different from the reference text, however, the concept is correct and clearly explained, being a better text than the similar text. Critically, the first text is going to score a higher BLEU and ROUGE than the second text, which is precisely what we do not want to happen.
Similar but wrong text: “Poisson Regression models are best suited for modeling events where the outcomes are counts, specifically continuous data with non-negative integer values following a normal distribution, like the number of times an event occurs within a specific period or the number of people in line at a store.”
In particular, we have computed the BLEU score for this text with the following value: BLEU = 0.836.
Different but correct text: “Poisson Regression models are ideal for predicting events represented by count variables. This type of data involves discrete, non-negative integers

that quantify occurrences, such as the frequency of an event within a specific period or the number of customers queuing at a store."
In this case, the BLEU score is 0.117, which is much lower than the previous case. This example illustrates how these quantitative measures, although they appeal ideal for this problem, will not measure the quality of the text generated, because we are not interested on replicating the reference text but in generating a new way to correctly express the underlying idea. If the text is almost similar but the concept is wrong, then, the text is wrong. However, as we have seen, BLEU and ROUGE will score a high value. Consequently, we rely on qualitative evaluations done by the experts, that give us a better estimate of the quality of the text, as BLEU and ROUGE are not statistics necessarily correlated with the quality of the text being generated.

**The statistical results indicate no significant difference between the outcomes produced by the GPT and BSVP methods. I recommend that the authors provide a more in-depth analysis of the interpretation of their results, particularly regarding the insights provided by the p-value.**
The explanation of the p-value has been expanded, including the effect size (Cohen's D) for each case in table 1. This explicitly demonstrates that the effect is negligible in all instances: it is below the 0.2 threshold typically considered indicative of a small effect. The following text has been added:
*the p-values obtained across the three dimensions, all of which exceed 0.05, suggest that the observed differences may simply be due to random variations in the data sample rather than a systematic effect of the customization implemented in the BSVP. Additionally, to complement these results, the effect size (Cohen's d) has been calculated for each dimension (effsize). Once again, a negligible effect size is confirmed in all cases, significantly below the threshold of 0.2 typically used to denote small effects. These effect sizes reinforce the conclusion that the customization of the BSVP system has not resulted in significant improvements in student interaction compared to the standard ChatGPT-4 Turbo model.*

**In my opinion, addressing these comments would significantly improve the quality of this research, enhancing its suitability for indexing.**
We sincerely appreciate the referee's comments.

***Competing Interests:*** No competing interests were disclosed.