Preliminary Evidence on the Use of Chatbots to Support Junior Software Professionals in Process Guidance

Fernando Rafael Serra Neves[1], João Victor Machado de Andrade[1], Julio Cezar Costa Furtado[1]

[1]Departamento de Ciências Exatas e Tecnológicas - DCET – Universidade Federal do Amapá (UNIFAP)
Rodovia JK, Km 02 – 68.902-280 – Macapá – AP – Brazil

{frsn777, joaovictormachadodeandrade}@gmail.com, furtado@unifap.br

Abstract.*This study explores whether AI-powered chatbots can enhance task performance and information access for junior professionals. A controlled experiment compared two groups: one used a chatbot, the other a wiki with expert support. Participants completed three tasks of varying difficulty and answered a feedback questionnaire. Preliminary results show that the chatbot group performed better, with statistically significant differences. They also perceived the chatbot as more accessible and efficient, despite noting issues with answer contextualization. These findings suggest chatbots are promising tools for onboarding and learning, though improvements are needed to handle more complex scenarios effectively.*

1. Introduction

The increasing complexity of software processes and the need for professionals to learn quickly require efficient solutions for information retrieval and task execution. Junior professionals, when they join development teams, are likely to face difficulties in learning complex workflows, which can impact their productivity and delivery quality (SANTOS et. al, 2024). Traditional methods of consultation, such as extensive documentation and the services of veteran experts, do not always suffice to provide information quickly and conveniently, hence necessitating research into other modes optimize this process.

Despite exposure to technical documentation and support from more veteran experts, most newcomers find it difficult to access extensive material and get straight answers to their questions (SANTOS et. al, 2024). This condition can create rework issues, increased time to deliver, and inconsistency in the use of processes. As artificial intelligence evolved, chatbot-powered virtual assistants come across as the perfect answer to make processes more streamlined for information retrieval and provide automated support, further increasing dependence on manual inquiries and improving professionals' productivity during training (SOARES & SILVA, 2024). However, few empirical facts remain concerning the effectiveness of chatbots compared to traditional ways of supporting software processes.

In this work, a controlled experiment was performed to test whether the impact of a chatbot on information recovery was comparable with that of a wiki and assistance from an experienced practitioner. A two-group trial utilized two sets of tasks under querying software workflows, and comparison of responses quantitatively applied performance metrics as well as statistical testing, combined with use of a feedback survey to gain impression of methods adopted by participants.

2. The Experimental Evaluation

The experiment was conducted to compare the effectiveness of different consultation methods in the context of an internal organizational development process, regarding new employees who need to gain knowledge about the workflow and business of the company. The study objectives were defined following the Goal, Question, Metric – GQM (BASILI, CALDIERA & ROMBACH, 1994):

- Study Objective 1: To contrast the learning effectiveness of junior professionals using a chatbot as a means of software process consultation with institutional wiki consultation and mentoring by a professional expert.
- Study Objective 2: To contrast the effect the chatbot-based solution has on perceived usability, effectiveness and comfort on the subjects with wiki consultation and guidance from an experienced professional.

To achieve these objectives, the following research questions were identified:

- Research Question 1: Does the utilization of chatbot as a source of consulting software processes enhance task execution performance, as opposed to consulting a wiki and advice from an experienced practitioner?
- Research Question 2: In what ways do various approaches to consulting software processes affect junior professionals' perception of usability, effectiveness, and comfort in accomplishing tasks?

The study involved 18 Computer Science students, all professionals in the field, randomly assigned to two groups: Group A used a chatbot to retrieve information, while Group B used an institutional wiki with optional expert support. The experiment, conducted in the second half of 2024, lasted two hours and was divided into three phases: an introductory presentation, task completion, and a feedback questionnaire.

Participants completed three practical tasks of increasing difficulty, designed to simulate real software process activities: (1) describing a single process activity, (2) detailing a workflow between two linked activities, and (3) constructing a full workflow involving multiple steps. All tasks were scored using a predefined rubric to ensure consistent and objective evaluation.

To assess perceptions (Research Question 2), a 5-point Likert scale questionnaire was applied, capturing participants’ subjective experiences with their assigned consultation method. This aimed not to evaluate the correctness of their answers but their impressions of the support tools used. Groups worked separately to avoid interference and were limited to their assigned consultation method throughout the experiment.

The tasks were designed to reflect realistic scenarios that junior professionals might face in software development environments. By progressing in complexity, the experiment aimed to test not only the ability to retrieve information but also to apply it correctly to practical situations. This approach allowed for a more comprehensive assessment of each method's effectiveness in supporting task execution.

Additionally, the inclusion of a feedback instrument provided valuable insights into the usability and perceived utility of the support tools. Participants' opinions helped identify not only performance outcomes but also how intuitive, comfortable, and helpful each method was from the user's perspective, an essential aspect when evaluating tools intended to support onboarding and learning.

3. Preliminary Results

3.3.1 Research Question 1

The Shapiro and Wilk (1965) test of normality was used to confirm if the means achieved by the students of the Experimental and Control groups were normally distributed. When applied to the whole data set, the test of normality yielded a W value of 0.92 and a p-value of 0.12, which confirmed the presence of normality of data. Then, Levene's test of homoscedasticity was used to examine the homogeneity of variances of the Experimental and Control groups. The result of the test was a Levene statistic equal to 1.23 and a p-value of 0.28. As the obtained p-value is greater than the chosen significance level ($\alpha = 0.05$), the null hypothesis of equal variances could not be rejected. Hence, it is concluded that there is homoscedasticity between the two groups, i.e., the variances of the data are homogeneous. According to the obtained results, the Shapiro and Wilk test of normality indicated that Experimental and Control group means are normally distributed ($W = 0.92$; $p = 0.12$), while the Levene test of homoscedasticity indicated that the variances of the groups are homogeneous ($W = 1.23$; $p = 0.28$).

Therefore, the parametric test assumptions for the application of parametric tests were satisfied. As a result, the two-tailed Student's t-test was employed to determine the difference in the mean of the two groups because the aforementioned test is suitable for independent samples with normal distribution and equal variances. The use of the two-tailed version is accounted for by the fact that there was no directional hypothesis in advance, i.e., the purpose is to validate whether there is any difference between the means of the groups that is statistically significant, but not which group would be higher or lower. To test whether there are statistically significant differences between the Experimental and Control groups, the following statistical hypothesis was formulated:

- Null hypothesis (H_0): The group means are equal, i.e., $\mu_1 = \mu_2$.
- Alternative hypothesis (H_1): The means of the groups are not the same, i.e., $\mu_1 \neq \mu_2$.

The findings revealed that the Experimental Group (6.63) had a significantly higher mean performance compared to that of the Control Group

(3.70). Application of two-tailed Student's t-test yielded a statistic t (16) = 3.33 with p-value = 0.004, which means that the observed group difference is statistically significant at the 0.05 level.

As the calculated p-value is less than the chosen significance level, the null hypothesis (H_0) is rejected, and it is concluded that the difference in the group means is statistically significant. This finding indicates that the utilization of chatbot as a means of asking questions to the software process contributed positively towards the response quality of the participants in comparison with the conventional method of written record and inquiry with a professional expert. Furthermore, the Cohen's d effect size was 1.57, which is a large effect, based on the criteria outlined by (COHEN, 1988). The result bears witness to the size of the difference between groups, confirming that the chatbot-based method of consultation not only had a statistically significant difference in the performance of the users, but also one that was substantive in the implementation of the tasks. Besides, to obtain more details regarding the experiment, a descriptive analysis of the averages obtained was made, where the Experimental group got an average of 9.00 ± 1.00 in the easy question, 6.33 ± 2.74 in the medium question, 4.56 ± 3.64 in the hard question and 6.63 ± 2.14 in the general average.

The control group achieved a mean of 4.11 ± 1.25 in the easy question, 3.67 ± 1.80 in the medium question, 2.89 ± 2.57 in the hard question and 3.70 ± 1.54 in the overall mean. So, the mean gain of $\Delta = 2.93$ on the total mean of the Experimental group indicates better performance compared to the Control group, another indication of a potential positive influence of the method employed. The findings indicate that the chatbot-based method led to considerably better performance in the completion of the task than the conventional wiki-based method through the assistance of an expert individual.

The Experimental Group's overall mean (6.63 ± 2.14) was higher than that of the Control Group (3.70 ± 1.54), which suggests that the chatbot users had a better performance in task completion. The Control Group, although with smaller deviation (2.57), still did worse, which implies that the standard method might have posed accessibility and utilization difficulties of information necessary to solve the more complex problems.

3.3.2 Research Question 2

The results provided by the feedback questionnaire indicate that there are substantial differences between the Experimental Group participants and the Control Group participants, illustrating that the consultation method directly affected the process of accessing information and undertaking tasks. The main findings are considered in more detail below.

To establish the reliability of the questionnaire that was used in the experiment, Cronbach's Alpha (1951) was calculated, a statistic coefficient widely used for measuring the internal consistency of research instruments. Cronbach's Alpha verifies the degree of correlation between the items of a questionnaire, whether the questions are measuring the same thing in a consistent way. The analysis provided $\alpha = 0.81$, which demonstrates the good reliability of the instrument, according to criteria established by Cohen (1988). The value expresses that the contents of the questionnaire have high internal consistency, i.e., the interviewees presented a stable and coherent response pattern, providing more strength to the validity of data collected.

The feedback questionnaire revealed substantial differences between the Experimental Group (chatbot) and the Control Group (wiki + expert) regarding ease of use, efficiency, and overall experience. The chatbot was perceived as significantly more accessible and intuitive. Most participants in the Experimental Group rated it as "easy" or "very easy" to use, while the Control Group predominantly found their method "difficult", largely due to the need to manually search extensive documentation.

Effectiveness was also rated higher in the Experimental Group, where participants noted quicker doubt resolution and clearer answers. The AI-based method provided direct and personalized responses, contrasting with the Control Group’s experience of slow information retrieval and cognitive overload from navigating long, unstructured texts. Participants using the chatbot reported fewer difficulties in finding information and more consistent support during task execution.

The frequency of consultation was high in both groups, but chatbot users found the process faster and smoother. Most chatbot users felt the tool increased their efficiency, while Control Group responses varied more widely, including some reports of reduced productivity.

In terms of comfort, the Experimental Group expressed greater satisfaction and ease when interacting with the system, while the Control Group reported discomfort and frustration. Open-ended responses revealed that chatbot users struggled mainly with formulating precise queries, while Control Group participants found the documentation poorly organized and difficult to search.

These preliminary findings suggest that while chatbots may not yet fully replace traditional methods, especially for complex tasks, their use can significantly enhance the learning experience, particularly for straightforward information needs. However, their effectiveness still depends on user ability to craft effective queries, indicating the need for improved interaction design and support mechanisms.

4. Conclusion

This study presented a preliminary investigation into the effects of different consultation methods for software processes on the performance of junior professionals. A controlled experiment was conducted with two groups: the Experimental Group, which used a chatbot to retrieve process information, and the Control Group, which relied on written documentation and the option to consult a senior professional. Participants completed practical tasks based on a defined process and provided feedback on their experience.

The initial findings suggest that the chatbot-supported group performed better in terms of task scores and perceived ease of use and effectiveness.

Participants in this group reported clearer and more useful information, along with a more fluid and comfortable consultation experience. In contrast, participants in the Control Group struggled mainly with navigating the extensive documentation and did not make use of the expert consultation option, raising questions about how support resources are actually utilized in practice.

As these results are preliminary, further studies are needed to validate the findings across broader contexts and participant profiles. Future research should explore hybrid approaches that integrate chatbots with structured documentation and expert guidance. It is also important to investigate how chatbot interactions can be personalized according to users' experience levels, and whether long-term use supports knowledge retention and increased autonomy in problem-solving. These avenues can help assess the true potential of chatbot-based support in professional training and software process learning.

Artifacts Availability

All artifacts related to this study — including datasets, experimental materials, and analysis scripts — are publicly available for consultation and reuse at the following Zenodo link: https://zenodo.org/records/15258493. This initiative aligns with open science principles, promoting transparency, reproducibility, and collaborative advancement of knowledge.

References

Basili, V. R., Caldiera, G. and Rombach, H. D. (1994). Goal/question/metric approach. In *Encyclopedia of Software Engineering*, volume 1, pages 528–532. John Wiley & Sons.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences.. Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. In *Psychometrika*, volume 16, pages 297–334. Springer.

Santos, Í., Soares, J. R. and Silva, P. N. (2024). Software solutions for newcomers' onboarding in software projects: A systematic literature review. In *Information & Software Technology*, volume 177, page 107568. Elsevier.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). In *Biometrika*, volume 52, pages 591–609. Oxford University Press.

Soares, J. R. and Silva, P. N. (2024). Panorama da pesquisa sobre chatbots no Brasil. In *Biblos: Revista do Instituto de Ciências Humanas e da Informação*, volume 38, number 1, pages 199–218. Universidade Federal do Rio Grande.