ISSN: 2473-4901
v11 n6435

# A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution

April Villeda Roblero
av9401@uncw.edu

Yang Song
songy@uncw.edu

Judith Gebauer
gebauerj@uncw.edu

Yao Shi
shiy@uncw.edu

University of North Carolina Wilmington
Wilmington, NC U.S.

## Abstract

Large language models (LLMs) and the use of artificial intelligence (AI) are becoming increasingly integrated into everyday life. As the models continue their training and improvement, leading technology companies are competing to create the most advanced artificial assistant. This competition has produced a variety of large language models, each with different capabilities. While LLMs like ChatGPT, Claude, and Gemini have proven effective at handling everyday tasks, they often lack the domain-specific expertise necessary for more specialized consultations. This limitation makes it essential to integrate a targeted knowledge base when developing chatbots for specific domains. Therefore, this study aims to address the question: Is there a model that clearly outperforms the others, considering correctness and comprehensiveness? In the study, we conducted an experiment to investigate the performance of chatbots utilizing LLMs within the context of a custom, mid-scale knowledge base designed for a university located in the southeastern United States. Using a web-crawled knowledge base, we created chatbots across multiple platforms and LLMs, testing them against a set of predefined questions to evaluate correctness and comprehensiveness. The findings highlight the disparity in the capability of LLMs and offer practical guidance for their effective use.

**Keywords:** Chatbot, Large Language Models, LLMs, AI, Higher Education

ISSN: 2473-4901
v11 n6435

# A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution

*April Villeda Roblero, Yang Son, Judith Gebauer, Yao Shi*

## 1. INTRODUCTION

Generative artificial intelligence has developed rapidly in recent years. Large Language Models (LLMs) represent a breakthrough in AI technology. These models are changing how people acquire, understand, and use information. Popular examples include ChatGPT (Firat, 2023), Claude (Liu et al., 2024), and Gemini (Islam & Ahmed, 2024). LLMs possess powerful natural language processing and knowledge generation capabilities. This enhances the efficiency and convenience of obtaining information through the internet. In fields that require domain knowledge and understanding of the context/scenario, such as consultation, LLMs have demonstrated "remarkable potential" (Song et al., 2023). In education settings, LLMs are also "promising tools for open education" as they can provide customized and interactive assistance to students and thereby improve the independence and autonomy of the learners (Firat, 2023). "Domain-specific" LLMs (Zhang et al., 2024) can respond quickly to diverse student needs. They provide personalized advice and information based on massive datasets. This capability reduces the burden on traditional consultation, support, and feedback methods.

However, LLMs face significant reliability and accuracy challenges in practical applications. The training data for these models is complex, diverse, and potentially biased. This leads to several problems during the understanding and inference processes. LLMs may misunderstand queries or engage in erroneous reasoning. They can even generate fabricated information, a phenomenon sometimes called "hallucination" (Chang et al., 2024). These issues highlight the gap between LLMs' demonstrated potential and their current limitations in real-world settings.

This uncertainty creates potential risks for users. Many users rely on LLM recommendations when making important decisions. Poor model accuracy could lead to harmful outcomes. Therefore, researchers must examine the accuracy and reliability of LLMs in consultation contexts. Such exploration is essential for promoting safe applications across different domains. Understanding these limitations will help ensure healthy deployment of LLM technology.

This study systematically evaluates the performance of major LLMs in answering a common set of questions. We used web crawling technology to collect extensive information about an R2 university in the southeastern United States. This information served as our knowledge base. The evaluation included two types of questions. **Objective questions** focused on new student enrollment topics such as tuition fees, accommodation, and food services. **Subjective questions** used specific scenarios and personas to assess contextual understanding. We input this data into multiple LLM platforms to create chatbots. The platforms included both commercial and open-source models: ChatGPT, Claude, Gemini, Copilot, Llama, and DeepSeek. These chatbots were designed to simulate real-life question-answering scenarios using the university knowledge base.

We compared and analyzed answers from different LLMs to achieve three main objectives. First, we demonstrate our evaluation framework for LLMs. This framework uses a knowledge base and question set that includes both objective and subjective items. Second, we explore the impact of reliable knowledge bases versus online search capabilities in building domain-specific chatbots. We analyze how these components improve answer quality and identify their limitations. Third, we propose strategies for improving LLM applications in educational consulting based on our research results. These recommendations serve as a reference for practitioners and researchers in related fields.

## 2. BACKGROUND

Chatbots represent an important research area within natural language processing (NLP), which is a subfield of AI that uses machine learning to help computers interpret, manipulate, and comprehend human language. Early machine translation research, a subdomain of NLP, relied on specific evaluation metrics. BLEU (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005) were two common approaches. BLEU evaluates translation accuracy by calculating similarity between machine-generated translations and

ISSN: 2473-4901
v11 n6435

reference translations. METEOR improves upon BLEU by introducing an alignment algorithm. This algorithm better handles synonyms and word order differences when evaluating similarity between generated and reference translations. However, these metrics are ineffective for evaluating chatbots.

Mehri and Eskenazi (2020) proposed the FED measurement metric to address this limitation. FED measures fine-grained dialogue quality at two levels. It evaluates individual dialogue turns, defined as "a dialog context and a system response (from chatbot)" (Mehri & Eskenazi, 2020), and entire dialogues. FED achieved moderate to strong correlation with human judgments at both levels.

Chatbot technology has advanced significantly with the introduction of LLMs. End users now have higher requirements for chatbot performance. User focus has shifted from basic fluency to multi-dimensional evaluations. Current assessments examine reading comprehension, reasoning capability, mathematical skills, and other technical measures (AI et al., 2024).

Higher expectations for chatbot performance bring attention to the Turing Test. Alan Turing, often called the "father of computer science," first introduced this concept in 1950. The test was originally known as the imitation game (Oppy & Dowe, 2021). The game involves behavioral evaluations that assess whether a machine can imitate human conversation. The key question is whether machine responses become indistinguishable from human responses. Turing argued that if a computer's response seems indistinguishable from a human response, we should consider whether it qualifies as a thinking entity. This question remains relevant today. Some research claims that modern LLMs have "passed" the Turing Test. This study does not conduct the Turing Test directly. However, we draw from similar principles by analyzing chatbot responses. Our focus is on their ability to provide accurate and contextually appropriate answers.

Researchers use specific datasets to evaluate dialogue response quality. These evaluations test fluency, naturalness, and other aspects to determine whether chatbots reach human-level performance. Mendonça et al. introduced a new evaluation benchmark called SODA-EVAL (Mendonça et al., 2024). This dataset used more than 120,000 turn-level assessments for training across 10,000 conversations. The researchers conducted human validation and annotation tasks to confirm automatic annotation quality. The evaluation system used a rating scale from 1 to 5 to assess dialogue response quality. Several factors determined these ratings. Evaluators examined whether responses contradicted dialogue history information. They assessed whether models covered all relevant conversation information. The system also measured fluency and naturalness of dialogue responses. This included evaluating common sense, participation levels, and repetition in conversations. The research revealed important findings about GPT-4 performance. Although GPT-4 performed well in many aspects, it showed room for improvement in coherence and common-sense reasoning. This indicates that large language models have made significant progress in generating fluent and relevant responses. However, they still face challenges in simulating the complexity of human dialogue.

Researchers have successfully created domain-specific chatbots that support domain experts in reading and decision-making. This progress has stimulated research in domain-specific chatbot evaluation methods. Song et al. (2023) developed an evaluation framework for medical applications. They created a questionnaire with 21 questions and two clinical scenarios related to urolithiasis. The researchers tested four LLMs: Claude, Bard, ChatGPT-4, and Bing. Domain experts evaluated the model responses using multiple criteria. These included accuracy, comprehensiveness, legibility, human care, and clinical case analysis ability. Evaluations used a 5-point Likert scale for systematic assessment. The study found that Claude and GPT-4 were the top-performing LLMs across their evaluation metrics. However, the research had a limited scope. The evaluation focused only on clinical urolithiasis-related dialogues, which represent a relatively narrow domain.

Educational applications of LLMs have also received research attention. Hwang et al. (2023) used an AI-driven approach to create and assess multiple-choice questions in chemistry and biology. They evaluated question quality using Item Writing Flaw (IWF) criteria (Breakall et al., 2019). The researchers combined machine learning models with human assessments to verify question alignment with Bloom's Taxonomy. The study used the RoBERTa model to validate 120 generated questions. A domain expert with over 28 years of STEM education experience assessed 57 of these questions. The research found that GPT-3.5 can generate questions aligned with Bloom's Taxonomy levels. However, notable differences emerged between human and machine quality assessments. These

ISSN: 2473-4901
v11 n6435

findings suggest a discrepancy between machine learning models and human evaluations when assessing AI-generated content. The results indicate that carefully considered evaluation standards are necessary for assessing AI-generated materials. This highlights ongoing challenges in developing reliable automated evaluation methods for educational content.

## 3. EXPERIMENT DESIGN

The scope of the experiment involves building chatbots on different platforms using data related to UNC Wilmington, a university we selected for this research, which is also the home institution of the authors. The data is crawled mainly from the website of the university under the uncw.edu domain. We also compare the performance of the chatbots when they answer evaluation questions with or without restricted internet access. In the data acquisition process, a web crawler was used to gather content from multiple pages within the domain from a URL such as https://library.uncw.edu/, https://uncw.edu/seahawk-life/dining-housing/housing/, or https://uncw.edu/research/. Those are the URLs that students may need to visit frequently, especially in their freshman year. The web crawling process ensures that the pages cover comprehensive information about this university, including details on housing, transportation, safety reports, meal plans, tuition fees, scholarships, and student lifeThe web crawler organized content into 20 topic-specific files. These files were saved in .docx, .pdf, or .csv formats. The collected data generated more than 148,000 tokens, which formed the foundation for the chatbot knowledge base. We conducted a preliminary comparison to determine the optimal file format. Two chatbots used the same knowledge base content, but one received a .pdf file while the other received a .docx file. The chatbot that used the .docx knowledge base performed better than its counterpart. Based on this finding, all subsequent chatbots in this study were trained using .docx files as their knowledge base unless otherwise noted.

We explored various open-source LLM and commercial AI chatbot frameworks to identify suitable tools for testing and comparison. Six platforms were selected for this project: GPT-4o, Gemini 1.5, Claude, Copilot Studio, Llama, and DeepSeek. We created multiple chatbots using these platforms and provided our knowledge base to each for testing. We included a baseline comparison by testing GPT-4o without any knowledge base. We also analyzed the impact of allowing chatbots to access the internet for searches in addition to using the knowledge base. For Llama, we tested different model sizes to assess performance variations. These included Llama 2 13B and Llama 2 70B models.

We focused exclusively on independent models and model providers when selecting commercial LLMs. Therefore, the "secondary" AI platforms that use those LLMs (such as BoodleBox) were excluded. **Table 2** in Appendix A lists the chatbots we have tested. When testing these LLMs, to minimize the impact of irrelevant factors, we provided the same prompt for the chatbots on each platform and conducted the tests using identical questions.

## 4. EVALUATION FRAMEWORK

We developed a structured framework to assess chatbot performance using two sets of testing questions. The first set contained 20 objective questions designed for single-turn interactions. In these tests, users asked one question and received one response without follow-up or context retention. This approach allowed us to evaluate basic chatbot performance. The second set introduced three fictional personas to test more complex interactions. Each persona engaged with chatbots through 13 subjective questions in multi-turn dialogues. These conversations involved multiple exchanges between users and chatbots. The format allowed for context retention and follow-up questions within coherent conversations.

| Objective Questions | |
|---|---|
| Q1 | What is the size of the school? |
| Q2 | What is the location of the school, rural or urban? |
| Q3 | What is the student-to-faculty ratio, and how large are the class sizes? |
| Q4 | What is the tuition cost, and what financial aid options are available? |
| Q5 | Does the school offer a computer science major? |
| Q6 | What is the college's retention rate, and how many students complete their degree |
| Q7 | How many students are there? |
| Q8 | What is the university's overall ranking among national universities? |

ISSN: 2473-4901
v11 n6435

| | |
|---|---|
| ow many residence halls or dorms does the school have? | |
| ow many dining halls or cafeterias are available for students on campus? | |
| hat percentage of incoming freshmen receive scholarships? | |
| hat is the reported campus crime rate per year by the school police department? | |
| hat is the average starting salary for students who graduate with a Bachelor's degree? | |
| ow many student organizations and clubs does the school have? | |
| ow to get access to the college's wifi? | |
| hat is the most recent year's acceptance rate for incoming freshmen? | |
| hat is the volume of the university's library collection? | |
| hat is the proportion of Asian students in the total number of students? | |
| hat are the school's wireless network coverage and stability metrics? | |
| ow many free or low-cost campus transportation options does the school offer? | |
| hich majors of this school are better known? | |
| oes the school offer any job connections with any company? [you may not find a good nswer] | |
| **Subjective Questions** | |
| hat is my intended major, and does the college offer a strong program for it? | |
| onsidering my preference of the campus size, does this university's campus size a good fit or me? | |
| ill I enjoy the location of this university considering my personal preferences? | |
| re there any on-campus housing options that will fit my lifestyle and preferences? | |
| hat on-campus clubs or activities can I join during my college years? | |
| ould the tuition be too expensive for me considering the tuition range I am okay with? nd considering my academic level, if there are possible scholarships for me from this niversity? | |
| ould you compare my intended major at this university with other university at similar evel? | |

| | | |
|---|---|---|
| | Q8 | How is the safety of the school compared to the average of the universities in the |
| | Q9 | What food options that aligns to my state does the on-campus dining offer? |
| | Q10 | Is there any famous scenery/place to visit outside campus that are aligned to my preference? |
| | Q11 | What kinds of sports team I likely to watch on campus at this university? |
| | Q12 | If I am taking a full student loan and using a 10-year payment plan, could you giv estimate about how much I'm paying after college? And is this too high for the ave salary of the students who graduate with my major? |
| | Q13 | Is it easy to rent an off-campus apartment near the school that is aligned to my pe need, and do I need to buy a car if I live off campus? |

**Table 3** in Appendix B lists the objective and subjective questions we created.

We created two test sets covering various aspects of university life. The first set contained 22 **objective** questions that focused on factual information. These questions addressed topics such as tuition and fees, housing options, and campus safety policies. The second set included three fictional personas and 13 **subjective** questions related to personal preferences. Example questions included "What food options that align with my taste does the on-campus dining offer?" and "Are there any on-campus housing options that will fit my lifestyle and preferences?" We developed detailed profiles for three virtual incoming college students to create these personas. Each profile specified preferred majors, food preferences, personal hobbies, and other individualized characteristics. This approach allowed us to test how well chatbots could provide personalized responses based on specific user needs and preferences.

We used different testing approaches for each question set. For the first set of questions, we conducted single-turn conversations to obtain and evaluate responses directly. This provided straightforward assessment of factual question handling. For the second set, we employed a multi-turn dialogue approach. We first provided each fictional persona's background information to establish context for the interaction. We then asked subjective questions from the test set. We expected chatbots to respond using both the knowledge base and specific persona details. **Table 3** in Appendix B lists all the assessment questions we have created for our experiments.

We evaluated chatbot dialogues using two key aspects: *correctness* and *comprehensiveness*. Each chatbot response received ratings on both

ISSN: 2473-4901
v11 n6435

dimensions using a 3-point Likert scale. Two researchers conducted the rating process independently. The inter-rater reliability achieved a Kappa index of 0.68, indicating substantial agreement between evaluators. The Likert scales were defined as follows in **Table 1**.

Please note that we used a 3-point Likert scale instead of the more commonly used 5-point version. We adopted simplicity and faster training of raters for the project timeline.

| Correctness | |
|---|---|
| 0 (Poor) | Wrong information (hallucination) or refusal to answer when the knowledge base includes the information needed. |
| 1 (Fair) | Partially wrong/inaccurate or misleading information provided, or refusal to answer when the knowledge base does not provide adequate information to answer. |
| 2 (Good) | Accurate information, or the information is very close to the information provided (considering some numbers, like student population, will change every year). |
| Comprehensiveness | |
| 0 (Poor) | Not enough information is provided to cover all the aspects of the question. |
| 1 (Fair) | Some information related to the question but inadequate, or too much additional and unrelated information. |
| 2 (Good) | Good coverage of all the aspects, no or limited unrelated information provided. |

**Table 1: Evaluation Scales**

## 5. EVALUATION RESULTS

**Objective Questions**

We created three ChatGPT-based chatbots with different configurations: GPT-4o without a knowledge base (GPT-4o), GPT-4o with knowledge base only (GPT-4o+KB), and GPT-4o with knowledge base plus search capabilities (GPT-4o+KB+S). The baseline GPT-4o relied on built-in knowledge and real-time internet searches. We found that GPT-4o was able to produce reasonably accurate and comprehensive answers that are comparable to GPT-4o+KB and better than Gemini, Copilot Studio, Llama 13B, Llama 70B, and DeepSeek. However, GPT-4o may make mistakes because the information searched may not be from the university website/domain.

The ChatGPT-based model that uses the given knowledge base but no real-time searches, GPT-4o+KB, could produce a similar level of correctness, but we observed that its performance of comprehensiveness was slightly weaker. The ChatGPT-based model that uses both the given knowledge base and online searches (the prompt required it to use the knowledge base as the *primary* information source), GPT4o+KB+S, produced the highest correctness and comprehensiveness among all the models. The results are listed in Appendix C (**Table 4**, **Table 5**)

Claude achieved the highest accuracy despite being an offline LLM without real-time search capabilities. Claude's correctness scores were phenomenal (tied for highest among all models). However, Claude frequently provided additional unrelated information, resulting in the fifth-highest comprehensiveness score due to irrelevant content. Claude's knowledge base capacity limitations required us to split our files and create four separate chatbots to answer all test questions.

Our Gemini-based chatbot missed important information from the knowledge base and often provided incomplete answers. For example, when asked about tuition, it provided only in-state rates while omitting out-of-state costs. Additionally, Gemini could not process tabular data such as Excel or CSV files. This limitation prevented it from accessing major lists and core course requirements, resulting in several "n/a" responses in our results.

The Copilot Studio-based chatbot received the most zero scores, indicating frequent inability to provide answers even when information existed in the knowledge base. Copilot provided shorter, less comprehensive responses compared to other chatbots. Despite allowing searches across four hyperlink domains, it performed weakest in both correctness and comprehensiveness. However, Copilot Studio offered the most extensive customization features, including conversational flow management, suggesting potential for handling complex tasks in future applications.

We tested two different Llama chatbots: Llama 2 13B and Llama 2 70B, both tested on LM Studio. Although Llama 13B is significantly smaller in size, it outperformed Llama 70B in both correctness and comprehensiveness. This suggests that larger model size does not necessarily lead to better performance, especially in domain-specific tasks. Llama70B struggled more with hallucinations but generally stayed on

topic. In addition, both models outperformed Copilot Studio.

DeepSeek 7B performed surprisingly well despite being one of the smallest models in the study. It received second place for comprehensiveness and third place for accuracy. DeepSeek frequently produced well-rounded responses, but it primarily suffered from providing partially incorrect information. Unlike Claude, DeepSeek was thorough enough without going into too much detail or overburdening the response. It is worth noting that DeepSeek is a relatively new model and was added towards the end of the study. Testing was restricted to a shorter period of time after access was made available via LM Studio. Despite these challenges, DeepSeek still performed well considering it's the smallest model tested in this study. **Tables 4 and 5** list the scores we have given to our chatbot on correctness and comprehensiveness.

**Subjective Questions**

We tested chatbot performance using three fictional personas with different university fit levels. The first persona was a poor fit based on intended major, location preferences, and city setting. The second was clearly a good fit considering academic plans and personal hobbies. The third represented a "borderline fit" for academic goals. These personas tested chatbot ability to answer subjective questions tailored to specific needs and interests. The results are listed in Appendix C (**Table 6**, **Table 7**, **Table 8**).

ChatGPT-based models (GPT-4o+KB and GPT-4o+KB+S) showed good but unstable performance across the three personas. Zero ratings primarily resulted from hallucinations where models fabricated information not in the knowledge base. For example, GPT-4o+KB+S incorrectly suggested a "Bachelor of Science in Civil Engineering" degree that wasn't on the provided major list. GPT-4o+KB, which relied solely on the knowledge base, showed fewer hallucinations.

Claude was most straightforward in advising "this university is not a good fit" for the first and third personas, making it the top performer for the first persona. However, Claude sometimes couldn't provide personalized suggestions due to limited information access (no real-time search capability). When lacking specific knowledge base information, it offered general responses instead. For example, when asked about off-campus apartment rentals aligned with personal needs, Claude provided only general information about housing availability and costs rather than personalized recommendations.

Gemini often ignored specific persona attributes when answering questions. For example, it calculated student loans using in-state tuition rates for an out-of-state student. Gemini provided lengthy responses with factual information that wasn't necessarily relevant to individual needs. However, it performed better than GPT-4o and Claude for the third persona. Gemini's inability to process tabular data prevented us from using all test questions, resulting in "n/a" entries in **Tables 6-8**.

Copilot Studio provided answers consistent with the knowledge base but often ignored fictional persona needs. For example, when the first persona preferred "her own apartment/suite with a private bedroom and bathroom," Copilot still suggested "double-occupancy, pod-style rooms." Copilot's responses were also less comprehensive than other chatbots. When asked to estimate student loans, most chatbots provided calculations and numbers, while Copilot offered only generic advice: "You can estimate your total student loan payment by multiplying the annual loan amount by the number of years and adding any interest accrued." While technically correct, this response wasn't helpful.

Both versions of LLama performed moderately; LLama 13B slightly outperformed LLama 70B in both correctness and comprehensiveness. LLama 13B tended to produce more grounded and relevant responses, particularly when dealing with questions related to housing, campus life, and tuition estimates. In contrast, LLama 70B occasionally hallucinated program names or exaggerated student life details that were not included in the knowledge base. However, both models did better than expected, especially considering they operated entirely offline without real-time search. Llama 13B and Llama 70B hallucinated, but their responses were often aligned with the persona's background. For example, both models sometimes offered generic recommendations without fully incorporating the unique hobbies or goals of the student persona. Nonetheless, their mostly accurate responses allowed them to outperform Copilot Studio and, in some cases, Gemini.

DeepSeek 7B proved to be both precise and efficient in addressing the subjective needs of the personas. Deepseek responded with relevant answers that aligned well with the persona's background and avoided unrelated content. The third persona was notably difficult for DeepSeek to handle. DeepSeek's responses lacked context

ISSN: 2473-4901
v11 n6435

of this persona, and it frequently overexplained or hallucinated information such as degrees or on-campus student services. Additionally, DeepSeek occasionally lacked deeper personalization when compared to models like Claude or GPT-4o+KB+S. Overall, DeepSeek's performance was solid and commendable.

## 6. CONCLUSION

In this paper, we crawled available information from a university in the southeastern United States, UNC Wilmington. The data covered multiple aspects, including academics, admission, student life, research, and university athletics. The volume of this unstructured knowledge base was comparable to that of a small-to-medium corporation (Jordão & Novas, 2024). We also included the university's annual security report as a PDF file. Additionally, we incorporated a major list and core course list that were crawled as Excel files. We created a testing framework to assess chatbot capabilities using the given knowledge base. The framework targeted chatbot performance in answering both objective and subjective questions. Objective questions focused on factual information while subjective questions addressed personal preferences. All questions were designed from the perspective of potential incoming students.

Among the commercial LLM models tested, ChatGPT-4o-based models performed strongest in both correctness and comprehensiveness for objective questions. This finding aligns with Rydzewski's research (Rydzewski et al., 2024). However, no single model consistently achieved the highest performance on subjective questions across different fictional personas. ChatGPT-4o, Claude, DeepSeek, and Gemini each demonstrated superior capabilities for different fictional personas. The following section presents our qualitative findings for different commercial LLM models.

ChatGPT-4o: Providing a customized knowledge base and utilizing online search options tends to improve responses for objective questions. However, hallucination remains a significant concern with this configuration. Maintaining the knowledge base while disabling online search can reduce hallucination occurrences.

Claude: This model does not offer real-time search capabilities. The responses are generally more "faithful" to the provided knowledge base. However, Claude's allowed knowledge base size is smaller than ChatGPT's capacity. This limitation makes it challenging to build domain-specific models on Claude when the knowledge base exceeds the textual data volume of an R2 university.

Gemini: This model does not support real-time searches. It typically excels at providing general answers and considering personal needs. However, the current version does not support structured input formats such as Excel files as part of the knowledge base.

Copilot Studio: This platform currently shows weakness in both correctness and comprehensiveness as of the first quarter of 2025. It is particularly weak in answering subjective questions related to specific persona needs. However, the developer's ambitious approach is evident. We remain hopeful for improved performance in future versions.

Llama: Both open-source models do not support real-time searches. The responses are stronger in relating back to the persona, but hallucination is of significant concern, especially with the 70B model. Despite the model being considered "open source," the limitations of licensing and regulations made the model difficult to use locally. The models performed relatively well overall and performed adequately on the subjective scenarios.

DeepSeek: The new model has generated significant buzz since its release, often compared to ChatGPT and its ability to compete. Although we tested a small 7B version, DeepSeek scored third place in the objective questions. Similarly to ChatGPT, it struggled with hallucination to a moderate extent. However, as the smallest model tested in this study, DeepSeek performed especially well in comprehensiveness. There is strong potential in this model as it continues to grow.

**Limitations**

The knowledge base we have created for this paper is for one single university, and yet this knowledge base cannot be processed by some commercialized LLM models (such as Claude). For even large knowledge bases, our approach will not be applicable.

This study aimed to create an assessment framework and evaluate current state-of-the-art LLM models. However, the rapid pace of innovation in this field will inevitably limit the generalizability of our results and observations. Our assessment framework has longer-lasting impact compared to our specific findings. The experiment was conducted between May 2024

ISSN: 2473-4901
v11 n6435

and March 2025. The new OpenAI o1 model was not included in the scope of this project.

Institutional or corporate regulations may prevent knowledge bases from being uploaded or shared online in certain cases. The technical approaches discussed in this paper will not be applicable under these circumstances. For researchers and working professionals facing such constraints, the optimal solution involves creating an in-house LLM server. This approach would utilize open-source LLM models such as Llama or DeepSeek.

**Future work**

For future work, instead of general daily tasks (like a college-related information chatbot in this paper), the authors are developing similar testing frameworks on more domain-specific chatbots with an unstructured or semi-structured knowledge base, like network system risk analysis.

## 7. REFERENCES

AI, et al. (2024). Yi: Open foundation models by 01.AI. arXiv. https://arxiv.org/abs/2403.04652

Banerjee, S., & Lavie, A. (2005). An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (pp. 65–72).

Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. Chemistry Education Research and Practice, 20(2), 369–382.

Chang, Y., et al. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45. https://doi.org/10.1145/3641289

Firat, M. (2023). How Chat GPT can transform autodidactic experiences and open education? OSF. https://doi.org/10.31219/osf.io/9ge8m

Hwang, K., Challagundla, S., Alomair, M., Chen, L. K., & Choa, F. S. (2023). Towards AI-assisted multiple choice question generation and quality evaluation at scale: Aligning with Bloom's taxonomy. In Workshop on Generative AI for Education. https://gaied.org/neurips2023/files/17/17_paper.pdf

Islam, R., & Ahmed, I. (2024). Gemini—the most powerful LLM: Myth or truth. In 2024 5th Information Communication Technologies Conference (ICTC) (pp. 303–308). IEEE. https://ieeexplore.ieee.org/abstract/document/10602253/

Jordão, R. V. D., & Novas, J. C. (2024). Information and knowledge management, intellectual capital, and sustainable growth in networked small and medium enterprises. Journal of the Knowledge Economy, 15(1), 563–595. https://doi.org/10.1007/s13132-022-01043-5

Liu, X., et al. (2024). Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: Comparative performance analysis. JMIR Medical Informatics, 12, e59273.

Mehri, S., & Eskenazi, M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT. arXiv. https://arxiv.org/abs/2006.12719

Mendonça, J., Trancoso, I., & Lavie, A. (2024). Soda-Eval: Open-domain dialogue evaluation in the age of LLMs. arXiv. https://doi.org/10.48550/arXiv.2408.10902

Oppy, G., & Dowe, D. (2021). The Turing test. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/turing-test/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). https://aclanthology.org/P02-1040.pdf

Rydzewski, N. R., et al. (2024). Comparative evaluation of LLMs in clinical oncology. NEJM AI, 1(5). https://doi.org/10.1056/AIoa2300151

Song, H., et al. (2023). Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. Journal of Medical Systems, 47(1), 125. https://doi.org/10.1007/s10916-023-02021-3

Zhang, Q., et al. (2024). A critical review of large language model on software engineering: An example from ChatGPT and automated

ISSN: 2473-4901
v11 n6435

program repair. arXiv. https://arxiv.org/abs/2310.08879

**APPENDIX A**
**LLMs Studied in the Project**

| Chatbot | Domain Based Chatbot Function with Knowledge Base Support | Cost | Search Online in Real-Time? | Additional Constraints |
|---|---|---|---|---|
| ChatGPT | My GPTs | $20/month | yes | Up to 20 files as the knowledge base |
| Gemini | Vertex AI | Free (Gemini 1.5 flash) | no | All the input knowledge base should be in one category: (1) PDF or doc, (2) JSON, or (3) CSV; 10 files, 1-million-line window size |
| Claude | Claude Pro projects | $20/month | no | Input knowledge base total line limit (20,000-40,000) because of Claude's maximum context window size |
| Copilot | Copilot studio | $200/month | Yes (but only from 4 links) | Can only support four URLs, two Dataverses (up to 15 tables), supports file uploads |
| Llama | LM studio | free | no | Supports file-based knowledge base |
| DeepSeek | LM studio | free | no | Supports file-based knowledge base |

**Table 2: LLMs Studied in the Project**

ISSN: 2473-4901
v11 n6435

**APPENDIX B**
**Questions for Chatbots**

| Objective Questions | |
|---|---|
| Q1 | What is the size of the school? |
| Q2 | What is the location of the school, rural or urban? |
| Q3 | What is the student-to-faculty ratio, and how large are the class sizes? |
| Q4 | What is the tuition cost, and what financial aid options are available? |
| Q5 | Does the school offer a computer science major? |
| Q6 | What is the college's retention rate, and how many students complete their degrees? |
| Q7 | How many students are there? |
| Q8 | What is the university's overall ranking among national universities? |
| Q9 | How many residence halls or dorms does the school have? |
| Q10 | How many dining halls or cafeterias are available for students on campus? |
| Q11 | What percentage of incoming freshmen receive scholarships? |
| Q12 | What is the reported campus crime rate per year by the school's police department? |
| Q13 | What is the average starting salary for students who graduate with a Bachelor’s degree? |
| Q14 | How many student organizations and clubs does the school have? |
| Q15 | How to get access to the college’s wifi? |
| Q16 | What is the most recent year's acceptance rate for incoming freshmen? |
| Q17 | What is the volume of the university's library collection? |
| Q18 | What is the proportion of Asian students in the total number of students? |
| Q19 | What are the school's wireless network coverage and stability metrics? |
| Q20 | How many free or low-cost campus transportation options does the school offer? |

| | |
|---|---|
| Q21 | Which majors of this school are better known? |
| Q22 | Does the school offer any job connections with any company? [you may not find a good answer] |
| **Subjective Questions** | |
| Q1 | What is my intended major, and does the college offer a strong program for it? |
| Q2 | Considering my preference of the campus size, does this university's campus size a good fit for me? |
| Q3 | Will I enjoy the location of this university considering my personal preferences? |
| Q4 | Are there any on-campus housing options that will fit my lifestyle and preferences? |
| Q5 | What on-campus clubs or activities can I join during my college years? |
| Q6 | Would the tuition be too expensive for me considering the tuition range I am okay with? And considering my academic level, if there are possible scholarships for me from this university? |
| Q7 | Could you compare my intended major at this university with other university at similar level? |
| Q8 | How is the safety of the school compared to the average of the universities in the U.S.? |
| Q9 | What food options that aligns to my state does the on-campus dining offer? |
| Q10 | Is there any famous scenery/place to visit outside campus that are aligned to my personal preference? |
| Q11 | What kind of sports am I likely to watch on campus at this university? |
| Q12 | If I am taking a full student loan and using a 10-year payment plan, could you give me an estimate about how much I'm paying after college? And is this too high for the average salary of the students who graduate with my major? |
| Q13 | Is it easy to rent an off-campus apartment near the school that is aligned to my personal need, and do I need to buy a car if I live off campus? |

**Table 3: Questions for Chatbots**

**APPENDIX C**
**Performance of Chatbots**

| Question | GPT-4o | GPT4o + KB | GPT4o + KB+S | Claude | Gemini - 1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
|---|---|---|---|---|---|---|---|---|---|
| Q1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 2 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 |
| Q4 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Q5 | 2 | 2 | 2 | 2 | n/a | 2 | 2 | 1 | 2 |
| Q6 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q8 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q9 | 0 | 1 | 2 | 2 | 1 | 0 | 1 | 1 | 0 |
| Q10 | 2 | 2 | 1 | 2 | 2 | 0 | 1 | 0 | 1 |
| Q11 | 2 | 2 | 2 | 2 | 0 | 0 | 1 | 1 | 1 |
| Q12 | 1 | 1 | 2 | 1 | 0 | 1 | 2 | 1 | 1 |
| Q13 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 0 |
| Q14 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q15 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| Q16 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q17 | 2 | 2 | 2 | 2 | 0 | 2 | 0 | 2 | 1 |
| Q18 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| Q19 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q20 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 |
| Q21 | 2 | 1 | 2 | 1 | 0 | 2 | 2 | 1 | 2 |
| **AVG.** | **1.67** | **1.67** | **1.90** | **1.90** | **1.50** | **1.33** | **1.48** | **1.38** | **1.52** |
| **Std. Dev.** | **0.58** | **0.58** | **0.30** | **0.30** | **0.83** | **0.86** | **0.68** | **0.67** | **0.68** |
| **CV.** | **0.35** | **0.35** | **0.16** | **0.16** | **0.55** | **0.64** | **0.46** | **0.48** | **0.45** |

**Table 4: Correctness of Chatbots on Objective Questions**

| Questions | GPT-4o | GPT4o + KB | GPT4o + KB+S | Claude | Gemini - 1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
|---|---|---|---|---|---|---|---|---|---|
| Q1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 0 | 2 |
| Q2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 2 |
| Q3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 0 | 0 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 2 | 1 | n/a | 2 | 2 | 2 | 2 |
| Q6 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 |
| Q7 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |
| Q8 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q9 | 0 | 1 | 2 | 2 | 1 | 0 | 1 | 1 | 1 |
| Q10 | 1 | 2 | 1 | 2 | 2 | 0 | 1 | 0 | 1 |
| Q11 | 2 | 2 | 2 | 1 | 0 | 0 | 2 | 2 | 2 |
| Q12 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| Q13 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| Q14 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| Q15 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| Q16 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q17 | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 2 | 2 |
| Q18 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |
| Q19 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q20 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |

| Q21 | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|
| **AVG.** | **1.52** | **1.48** | **1.86** | **1.52** | **1.30** | **1.33** | **1.67** | **1.57** | **1.76** |
| **Std. Dev.** | **0.68** | **0.68** | **0.36** | **0.60** | **0.73** | **0.73** | **0.48** | **0.68** | **0.44** |
| **CV.** | **0.45** | **0.46** | **0.19** | **0.39** | **0.56** | **0.55** | **0.29** | **0.43** | **0.25** |

**Table 5: Comprehensiveness of Chatbots on Objective Questions**

| | **Correctness** | | | | | | | | **Comprehensiveness** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Questions** | **GPT4o + KB** | **GPT4o + KB+S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** | **GPT4o+ KB** | **GPT4o+ KB+S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** |
| Q1 | 2 | 2 | 2 | n/a | 0 | 0 | 0 | 1 | 2 | 2 | 1 | n/a | 0 | 1 | 1 | 1 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 0 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| Q6 | 0 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 2 | 1 | 1 | 2 | 0 | 2 |
| Q7 | 0 | 0 | 2 | N/A | 0 | 0 | 0 | 0 | 1 | 1 | 2 | N/A | 1 | 1 | 1 | 1 |
| Q8 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q9 | 1 | 2 | 0 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q10 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| Q12 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q13 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 0 | 2 | 2 |
| **Avg.** | **1.38** | **1.54** | **1.85** | **1.82** | **1.23** | **1.23** | **1.15** | **1.54** | **1.46** | **1.69** | **1.69** | **1.45** | **1.15** | **1.54** | **1.46** | **1.77** |
| **Std. Dev.** | **0.77** | **0.78** | **0.55** | **0.40** | **0.73** | **0.73** | **0.69** | **0.66** | **0.66** | **0.48** | **0.48** | **0.52** | **0.55** | **0.66** | **0.78** | **0.44** |
| **CV.** | **0.55** | **0.50** | **0.30** | **0.22** | **0.59** | **0.59** | **0.60** | **0.43** | **0.45** | **0.28** | **0.28** | **0.36** | **0.48** | **0.43** | **0.53** | **0.25** |

**Table 6: Performance of Chatbots on Subjective Questions (First Persona)**

| | **Correctness** | | | | | | | | **Comprehensiveness** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Questions** | **GPT4o + KB** | **GPT4o + KB+S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** | **GPT4o+ KB** | **GPT4o+ KB+S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** |
| Q1 | 2 | 2 | 2 | n/a | 2 | 2 | 2 | 1 | 2 | 2 | 2 | n/a | 1 | 2 | 2 | 2 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 2 |
| Q6 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| Q7 | 1 | 2 | 1 | n/a | 1 | 1 | 1 | 1 | 0 | 2 | 2 | n/a | n/a | 1 | 2 | 1 |
| Q8 | 1 | 2 | 2 | 1 | n/a | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 |
| Q9 | 2 | 2 | 2 | 1 | 2 | n/a | 1 | 2 | 2 | 2 | 2 | 2 | 2 | n/a | 0 | 2 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q10 | 2 | 2 | 2 | 2 | n/a | 2 | 1 | 1 | 2 | 1 | 2 | 2 | n/a | 2 | 0 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 2 |
| Q12 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q13 | 2 | 2 | 2 | n/a | 1 | 2 | 2 | 2 | 2 | 2 | 1 | n/a | 1 | 2 | 2 | 1 |
| **Avg.** | **1.69** | **2** | **1.69** | **1.5** | **1.27** | **1.75** | **1.62** | **1.69** | **1.54** | **1.77** | **1.77** | **1.67** | **1.09** | **1.83** | **1.54** | **1.85** |
| **Std. Dev.** | **0.48** | **0.00** | **0.48** | **0.53** | **0.79** | **0.45** | **0.51** | **0.48** | **0.66** | **0.44** | **0.44** | **0.48** | **0.70** | **0.39** | **0.78** | **0.38** |
| **CV.** | **0.28** | **0.00** | **0.28** | **0.35** | **0.62** | **0.26** | **0.31** | **0.28** | **0.43** | **0.25** | **0.25** | **0.28** | **0.64** | **0.21** | **0.50** | **0.20** |

**Table 7: Performance of Chatbots on Subjective Questions (Second Persona)**

| | **Correctness** | | | | | | | | **Comprehensiveness** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Questions** | **GPT4o + KB** | **GPT4o + KB+ S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** | **GPT4o+ KB** | **GPT4o+ KB+ S** | **Claude** | **Gemini-1.5-flash** | **Copilot Studio** | **Llama 13B** | **Llama 70B** | **DeepSeek 7B** |
| Q1 | 1 | 1 | 2 | n/a | 0 | 0 | 0 | 0 | 2 | 2 | 2 | n/a | 1 | 1 | 1 | 0 |
| Q2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 2 |
| Q3 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 0 | 1 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| Q5 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| Q6 | 0 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 |
| Q7 | 2 | 0 | 1 | n/a | 0 | 0 | 0 | 0 | 2 | 1 | 2 | n/a | 0 | 1 | 1 | 1 |
| Q8 | 2 | 2 | 1 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |
| Q9 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| Q10 | 1 | 2 | 1 | 2 | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Q13 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 |
| **Avg.** | **1.54** | **1.46** | **1.54** | **1.72** | **1** | **1.15** | **1.15** | **1.30** | **1.85** | **1.62** | **1.62** | **1.73** | **0.85** | **1.38** | **1.62** | **1.54** |
| **Std. Dev.** | **0.66** | **0.66** | **0.52** | **0.47** | **0.82** | **0.69** | **0.69** | **0.95** | **0.38** | **0.51** | **0.51** | **0.47** | **0.55** | **0.65** | **0.65** | **0.66** |
| **CV.** | **0.43** | **0.45** | **0.34** | **0.27** | **0.82** | **0.60** | **0.60** | **0.72** | **0.20** | **0.31** | **0.31** | **0.27** | **0.66** | **0.47** | **0.40** | **0.43** |

**Table 8: Performance of Chatbots on Subjective Questions (Third Persona)**