An Overview of the use of Small Scale AI and LLM Models in the context of Receptionist Chatbots

MSc Research Project
MSCDAD JAN23A I

Talha Masood
Student ID: 20164467

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Talha Masood
Student ID:	20164467
Programme:	MSCDAD JAN23A I
Year:	2023
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	14/12/2023
Project Title:	An Overview of the use of Small Scale AI and LLM Models in the context of Receptionist Chatbots
Word Count:	5500
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Overview of the use of Small Scale AI and LLM Models in the context of Receptionist Chatbots

Talha Masood
20164467

Abstract

LLM (large language models) are becoming quite widespread in use. They have many use cases, especially in the form of specialized chatbots. One such use case is the use of such models as a receptionist chatbot in an office environment. There is a need to look at various aspects of the outputs of such LLM in order to see how their responses are even suited for use in such models or not. A variety of tools can be used to gauge how well the models perform this work. Using word clouds, sentiment analysis is a baseline way of looking at whether the responses are provided in a professional way or not. In addition, hand collected LLM conversations can be seperated into several groups depending on whether or not the request of the prompting user was adequately responded to. The user in this case would ask the chatbot the location of a virtual office environment and the chatbot would be expected to answer with some degree of accuracy. Overall results have been impressive and showcase the performance of existing models.

1 Introduction

1.1 Aims, background, motivation

LLM models are becoming very popular. Modern technologies move towards using them. There are many conceivable use cases. Hence, there is a need to see the possibility of such use case before implementation to prevent unnecessary damage.

LLM can be used to make chatbots which have a very long and well developed history (Suta, Lan, Wu, Mongkolnam and Chan 2020) and (Rudolph, Tan and Tan, 2023). A variety of purposes have been catered to, including social and entertainment purposes. This can also be extended to a job setting. A receptionist job is a simple use case to test this idea out. Inputs will be what is asked of the receptionist. The output is the response to the query.

1.2 Research question

The task is to guage how good an existing small scale model performs upon having to serve as a receptionist chatbot. The Research Question, is as follows:

”How effective and comprehensive would currently existing AI and small scale LLM Models be in implementing a fully functional receptionist chatbot that needs to fulfil the tasks required to satisfy customer queries?”

The aim is to answer if the tasks of a receptionist can be performed by current small scale AI LLM models. An example of such a task is to see if the user can adequately reach a target location or not.

1.3 Objectives

The next objective is to train several chatbots to perform as a receptionist in an office environment. The appropriate prompts will be used to accomplish this. Various models can be used including Vicuna, LlaMA and Wizard. The chatbots will be asked the directions to get to a certain person in the simulated office eninronment. The responses will be analysed to quantify how well the query was responded to. There will be three parts to this. 1) Wordcloud in order to see if the word choice is professional. 2) Sentiment analysis to see if the attitude was positive. 3) Seperation of responses into groups based on how accurate the directions given are.

1.4 Justification

This study can have widespread benefits. It can make it easier to verify if smaller LLM are enough when it comes to catering to customers. There would be lesser need to jump to larger and more costly solutions that cant be run locally.

1.5 Stakeholders

The actual end users interacting with the chatbot are the first stakeholders and they would mainly beskilled or nonskilled in terms of user group. The responses by the caht-bot could either leave them satsfied or non satisfied. A satisfied customer would benefit the second stakeholder who would be the management of the organization or building where the chatbot is implemented. A poorly made chatbot would mean less positive customer feedback in this regard.

1.6 Proposed Contributions

The main contribution would involve comparative analysis of existing LLM Models which are smaller in terms of parameters. There have been very few such studies. We can see how well existing models perform as well as see if some models perform better than others.

2 Related Work

To start with, we need to confirm that there have in fact been several LLM models released in revent years. Many of them have a variety of capabilities. (Touvron, Martin, Stone, Albert, Almahairi, Babaei, Bashlykov, Batra, Bhargava, Bhosale and Bikel, 2023). Various work into instruction tuning has also been done in regards to these language models (Peng, Li, He, Galley and Gao, 2023). These are very recent works and we can be assured that relevant work is ongoing. Even in past years, there have been many studies attempting to analyse AI chatbots before the LLM boom in the past few years (Lokman and Ameedeen, 2019).

One of the many studies that we can see in this regards has to do with a particular paper that was released in 2020. This paper takes research from between the years of 1998 to 2018 into account. A close glance into the paper shows that it looks into creating a mind map of sorts which is made with the intention to understand the different ways and concepts used in the creation of chatbots. Challenges and shortcomings of such implementations are one of the points discussed. The notion of context is very important in this paper and the paper discusses that for a good chatbot, it is neessary to have the right understanding of the context. It is brought into attention that many current implementations at the time of writing the paper are actually quite lacking when it comes to the notion of context and the ability to properly understand context. This was a useful study since the purpose of this research was also in tune with how well chatbots perform. By the end, we intend to make similar findings about the performance of existing models right now. There may be shortcomings that are being faced right now that were not present during the time of the previous papers. The paper in question is (Suta, Lan, Wu, Mongkolnam, and Chan, 2020).

The previous discussed paper is a good example of a study of chatbot performance and how it can be improved. However, it is not the only one we can consider. There are other papers that takes previous publications into account as well. One such paper focuses on research that was done between the years of 2017 and 2023. The paper dealt with the current landscape when it comes to LLM models that were published around these times. It takes a huge number of publications into consideration. This includes over 5000 papers. The papers were analysed using the "Topic Modelling". Using such a method, a summary can be provided from latent semantic topics of huge amounts of data which is unstructured. The paper in question is (Fan, Li, Ma, Lee, Yu, and Hemphill, 2023).

We can take the concept of literature metadata into consideration as well since this paper focuses on this too. In fact, a method called BERTopic was used in the paper mentioned earlier. This is described as a neural topic modeling method. Publications and texts can be analysed using it. As a result of using the method, around 200 topics were collected which were based on the keywords used in the research papers. A Web of Science Core Collection was used to collect the needed data.

There are many papers that deal with the implementation of chatbots as well. Various and numerous methodologies can be considered in such AI chatbots. A number of such methodologies were discussed which were used between the years 1999 to 2022 in one such research paper. The amount of text studied was quite large and around 3372 documents were used in order to make the discussion more broad and diverse. Many topics were discussed which included the reasons as to why such chatbots were built and what their objectives were. Methods in making such implementations were also discussed as well as what eventual outcome was encountered by the end of the implementation. The paper in question is (Lin, Huang, and Yang, 2023).

We can also at the same time have a look at research papers that take into account a lot of smaller parts that make up the whole of an LLM. In this sort of area, we can find a research paper which discusses many aspects of the topics such as findings, key techniques and backgrounds of these LLMs. This paper identifies the pre training , adaotation tuning, capacity evaluation and utilization of the LLMs in the process. The paper in question is (Zhao, Zhou, Li, Tang, Wang, Hou, Min, Zhang, Zhang, Dong, and Du, 2023).

We also need to consider research that takes future steps into account as well. One such paper does just that abd in addition, it also has a discussion regarding the various emergent abilities that LLM models have. This paper specifies that several new abilities have come into existence as a result of the developement of LLM models and their subsequent increase in scale. There is a question that arises in such a situation. We need to ask if increasing the size further will result in newer abilities. The scale of size used in this paper is huge and we will not discuss such large scale. We would instead focus on 7b and 13b sized models in particular. The paper in question is (Wei, Tay, Bommasani, Raffel, Zoph, Borgeaud, Yogatama, Bosma, Zhou, Metzler, and Chi, 2022).

Up to this point, the majority of papers that we have discussed so far are of a very specific variety. They are mostly based on surveys and existing literature reviews. We cannot just stop just there since that is not the only type of research that we are interested in. We also have to look at actual implementations of chatbots as well. This can come in handy when coming up with a sort of methodology that can make a functional chatbot in some shape or form. Then we can scrutinize it further to see if it really is a proper implementation that fulfils its purpose or not.

To start with, we can have a look into one such paper which takes a look into the topic of customer support. This is a sort of domain that is along the same lines as our topic. We also aim at creating a sort of receptionist chatbot for the purposes of this research. The paper in question is (Barbosa and Godoy, 2021).

There has previously been a great amount of work done in the past that does not entirely encompass the scope of completely integrated chatbots. One example that we can look at is in regards to the use of customer service agents which are used in smart dialogue systems. An notable example that we have of such a work is mentioned in a paper which touches upon discussions of implementing the resolution to issues that can result from shortage of context in conversations with clients. Doing so, we can get insight into the issues that can be faced while improving such dialogue systems. The paper in question is (Wan and Chen, 2018).

This paper uses the BERT model. This is a language model that is used in cases involving working with NLP. The chatbot implemented here is a receptionist chatbot.

The bot uses a finite state machine so that it can help clients of a real estate company. Such a paper demonstrates working models of such a chatbot and insight into how it can be implemented.

When we look at further papers, we can run into research that looks into the aplication of NLP and ML. This is used in order to create a customer service chatbot. This sort of chatbot is only effective if it is able to respond in the appropriate time. We cannot say that we have successfully made a workable implementation of the chatbot if the responses are not coming in a timely way. The paper that discusses this is (Liu, Jiang, Xiong, Yang and Ye, 2020). This paper discusses how to solve this by using a method called MRTM. MRTM stands for Multi-Turn Response Triggering Model. In the paper, we get to know that replying each and every time the user asks a question is not the most useful response as we may run into a situation where the answers are not even immediately helpful to the client at all. Misleading and incorrect answers are a huge possibility at this point. The proper context and ability to answer can be improved by using large scale human to human dialogues. According to the paper, this sort of effort makes the results much better.

A very interesting paper suggests using LLM itself to judge different chatbots (Zheng, Chiang, Sheng, Zhuang, Wu, Zhuang, Lin, Li, Li, Xing and Zhang, 2023). Although seperate from chatbots, LLM such as Vicuna has also been used to enhance contract drafting (Lam, Cheng and Yeong, 2023). We have also seen the use of LLM chatbots to create tourist recommendations (Arteaga, Arenas, Paz, Tupia and Bruzza, 2019). And although not used to simulate an office environment, we can see that an attempt at conceptualizing an LLM playing the role of a receptionist at a library in one paper (Bagchi, 2020). Another creative way envisioned to use LLM is in a learning environment where someone learning to read can have a conversation with an LLM regarding the content that has just been studied(for example, to help in assessment and feedback) (Wong, Lacey, Gharpure, Hao, Venkatraman, Elidan, Engelberg, R., Hackmon, Rabin, Fink and Yu, 2023).

Although there has been a lot of work done in recent years regarding LLM and AI, ecspecially in the context of chatbots, we still need a study that focuses on one particular use case and observes the responses given. A focused implementation and analysis has still not been performed. A side by side comparision of the performance of several chatbots when given the same task might be a way of improving this.

3 Methodology

The methodology involves several steps. A virtual office environment would be prompted to each LLM being tested. The prompt would include directions to each person in the office. Upon being asked, the chatbot must give the appropriate response. In the end the results will be tabulated according to their accuracy. A pie chart will be used to seperate the answers according to their categories.

The LLM models will be accessible through a service called Chatbot Arena. Hundreds

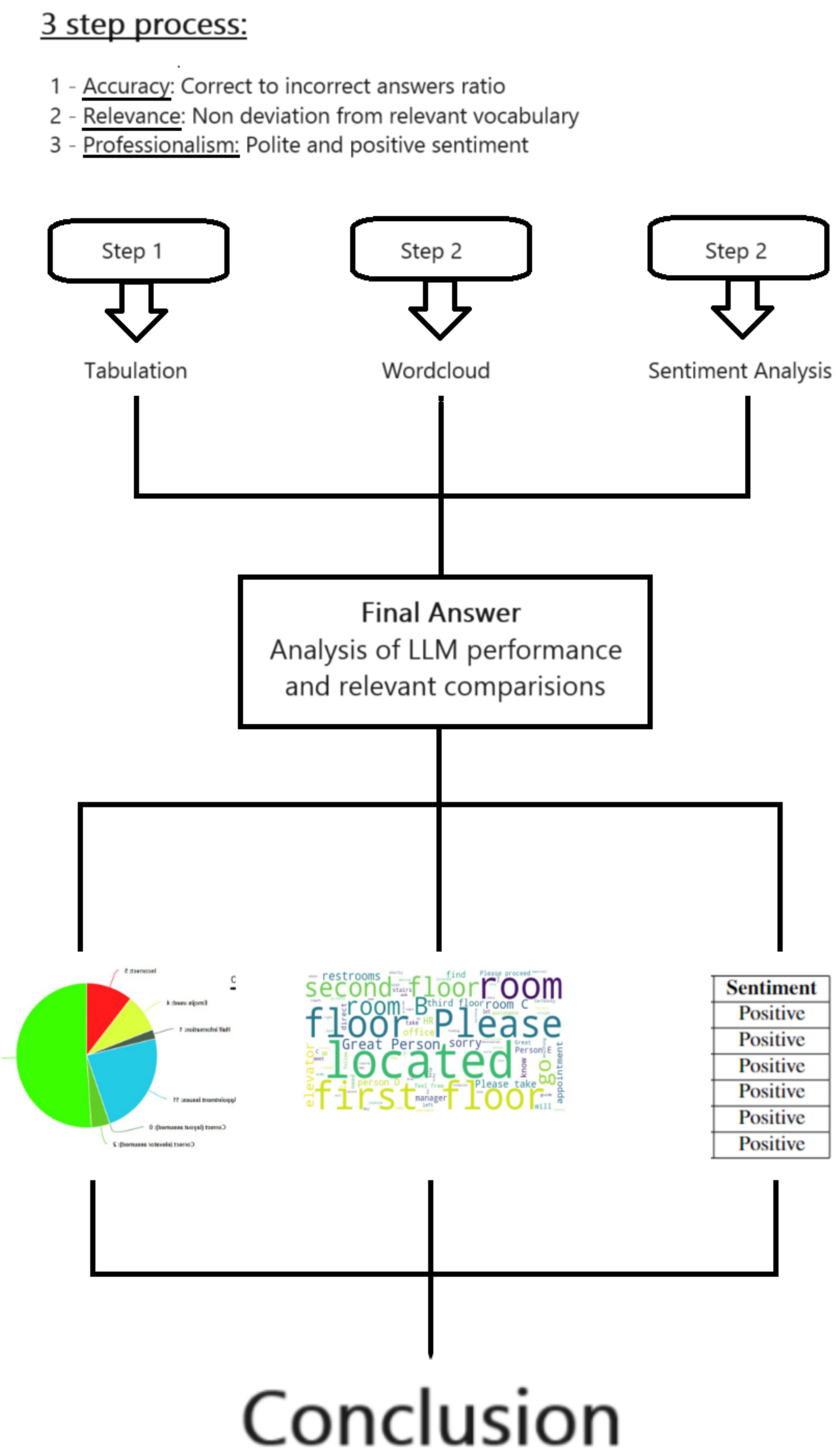


Figure 1: Process to be followed

of responses will be collected by hand and then seperated into various folders according to answer accuracy. Wordcloud and sentiment analysis will also be performed to guage the quality and proffesionalism of the answers.

We can imagine that the virtual environment in which our receptionist dwells is the form of an office complex with a variety of rooms. We can assume that a person seeking to find one of the rooms would come to the receptionist in order to inquire about it. The receptionist chatbot would then proceed to give an answer which would then lead the inquiring person to a specific location.

A successful encounter with the receptionist would mean that the location answered by the chatbot and the inquired location is the same. An unsuccessful encounter would mean that the chatbot would not answer correctly. In the case of large scale LLM models such as ChatGPT where you have many more parameters to take into account, this sort of conversation would be rather trivial.

There are 175 billion parameters to be considered here in GPT4 comparison to many of the other offerings available in case of smaller LLM models. 175 billion parameters is too much to be run on any sort of currently existing easily attainable consumer level hardware. Smaller models are much better suited to be run on local hardware (around 13b is an example of small scale in this context).

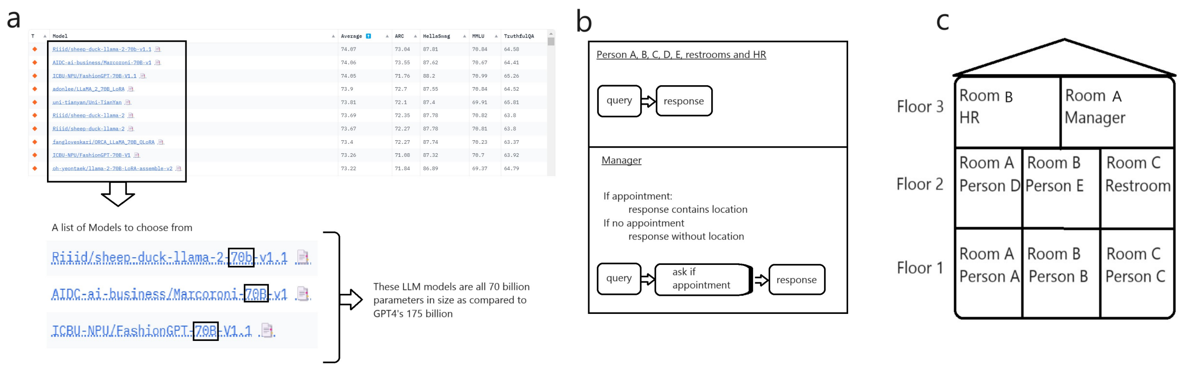


Figure 2: a) showing various LLM models with their sizes on the website "Chatbot Arena". b) Digram showing expected responses for user queries c) Virtual office environment being visualized

The virtual office environment is defined as follows: 1) The first floor would contain three rooms. Room A, Room B, and Room C. These would contain Person A, Person B, and Person C respectively. 2) The second floor would contain an additional three rooms. These are Room A, Room B and the restroom. Rooms A and B of the second floor would contain Person D and Person E. A third room would be be the restroom area. 3) On the third floor we have only two rooms. These are the HR room and the Managers room. In regards to every room other than the managers room, simply prompting the receptionist chatbot to tell the directions to the room would suffice and the chatbot would answer accordingly. However, in the case of the manager, the chatbot would first need to inquire if the prompter has an appointment or not. Directions to the managers office should only be provided if there is an appointment.

There are various LLM that can be used. A few of them are as follows: 1) LLaMA: This stands for Large Language Model Meta AI. The smallest variant of this model is the LLaMA 7b model. It is described to have been designed to help researchers advance their work and research in AI. It is described as requiring much smaller requirements to work with and discover new use cases. Large amounts of insfrastructure are not needed. Several sizes are available, but for the purposes of this task, only the 7b and 13b variants will be used since they are smaller in size by comparision to the others. 2) Vicuna: Vicuna is an open source LLM chatbot which claims to generate responses that rival that of ChatGPT. One of its versions, Vicuna 13b, was trained by fine tuning LLaMA on user shared conversations collected from ShareGPT. This results in a versatile LLM that is capable of generating detailed answers. The code, weights and demo are all entirely open source.

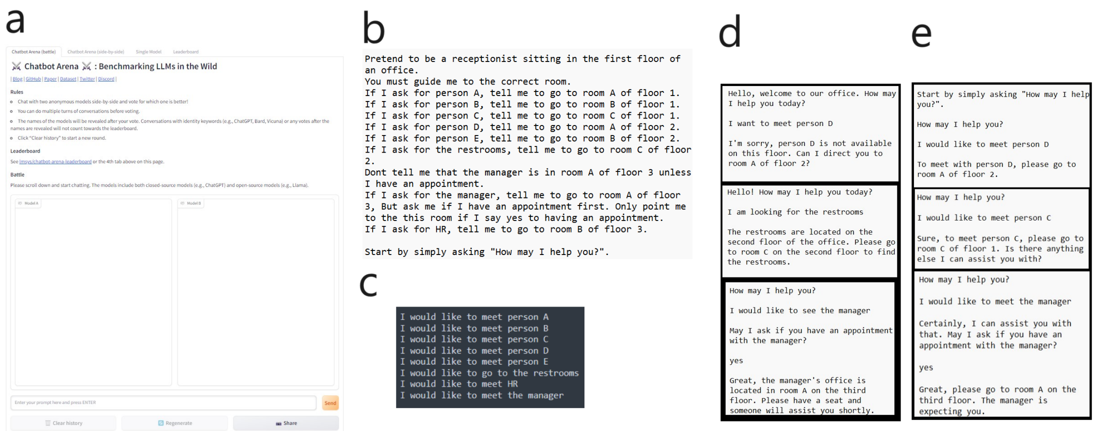


Figure 3: a) The Chatbot Arena website. b) The initial prompt given to each model. c) The subsequent requests that can be made by the user. d) Some sample conversations by the Vicuna 7b model (note that in the last conversation, the user is not given the directions to the manager unless there is a confirmed appointment). e) Some sample conversations by the Vicuna 13b model

We can observe that the responses provided by the various models can vary. There are some similarities, but in large, the different models can greatly vary in the sort of responses that we can acheive. It should be taken in great consideration whether the model being used is actually inductive to the sort of responses that we actually require.

We can analyse some common patterns. Firstly, we can see that the initial greeting may vary. It is however almost always very brief and assumes that the user just walked into the office building at the same time. Details like the elevator and stairs are automatically inferred by the chatbot. Further training of the bot with further prompts and commands might direct the chatbot to only include such details if the building in question has such amenities at all. The response at the end contains the location of the room being inquired about.

After looking at the several sorts of responses that can be given. We can make certain categories of answers provided by the LLM: 1) Correct, 2) Correct (elevator assumed) ,3) Correct (layout assumed) ,4) Appointment issues ,5) Half information ,6) Emojis used, and 7) Incorrect. Their descriptions are as follows:

1) Correct: This is a correct response. There are two categories that must be fulfilled for a response to be categorized as correct. The first one is the proper identification of the floor on which the target person is located. The second is the proper identification of the room in which the target person is located within that floor. This type of response also contains no "artifacts" or additional unrequired information. It will also not have any made up portion regarding which we have not given any information in the prompt. The made up information more often than not involves guiding the user along the stairs or elevator when there has been nothing in the input prompt that would signify that the office building being simulated even has an elevator to begin with. This sort of response is ideal for the user to have in response for the query.

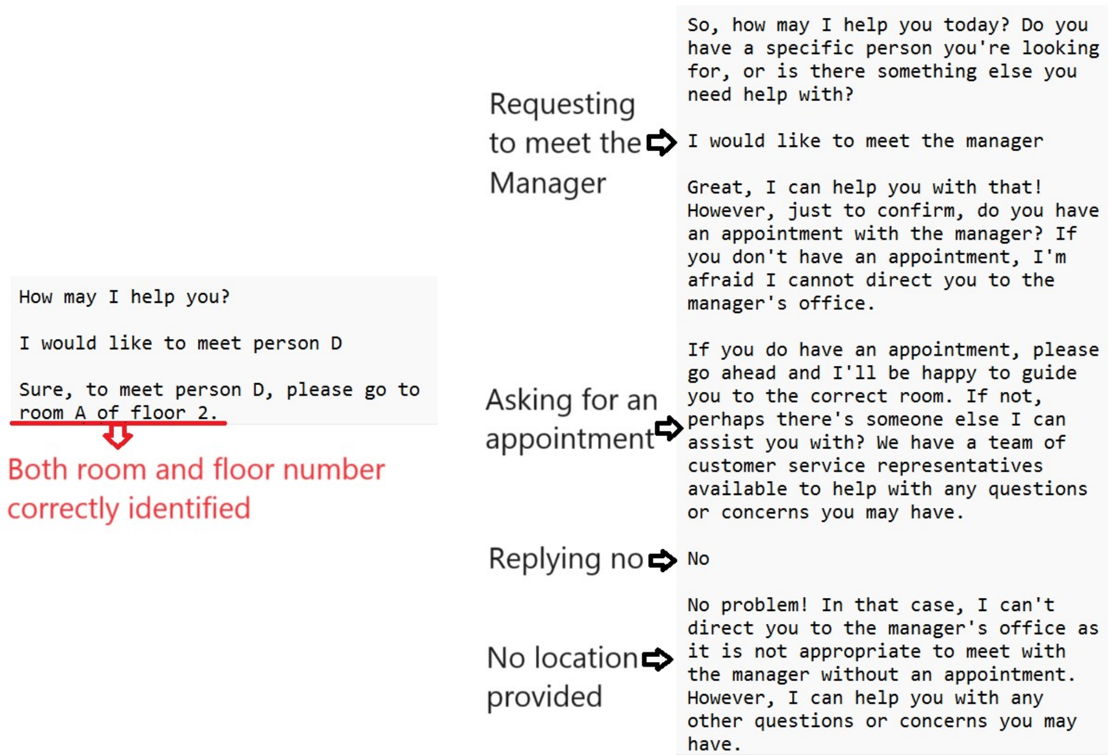


Figure 4: Two examples of correct responses. The location in terms of room and floor are correctly idetified in the left, and the location is not disclosed unless the appointment with the manager is confirmed in the right.

2) Correct, but elevator assumed: This sort of output is just as good as the correct outputs. It fulills the two criteria mentioned earlier without much problem. The first one is the proper identification of the floor on which the target person is located. The second is the proper identification of the room in which the target person is located within that floor. However, this response contains details regarding the presence of an "elevator"

within the building, which is not known by the LLM as of yet. A single look at our initial prompt signifies that we have not given the LLM any information that suggests that there is any sort of elevator within the building. This sort of response gives the correct location, but might mislead the user into looking for an elevator leading to the room even though the elevator itself does not exist.

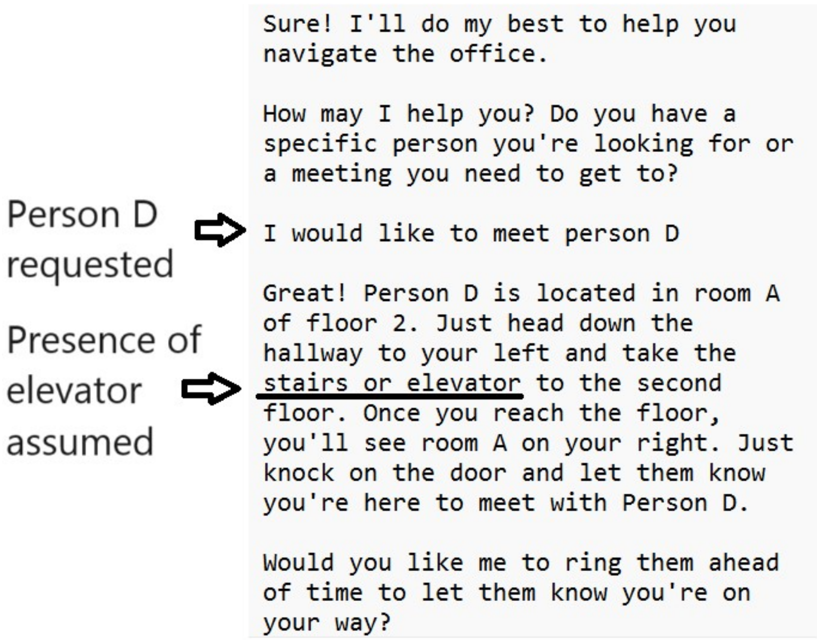


Figure 5: An example where the answer is correct, but the existence of the elevator is assumed

3) Correct, but building layout assumed: This response is sort of an extension of the problem as mentioned above. We have the correct answer. The two criteria mentioned would be met as before. The first one is the proper identification of the floor on which the target person is located. The second is the proper identification of the room in which the target person is located within that floor. However, instead of merely assuming the presence of an elevator, this type of response assumes the interior and layout of the office building. It might make an indication of there being a hallway, or the room being towards the users right or left, whereas the LLM model would have no way of knowing for sure if this is the actual layout of the building itself. Simply telling the room and floor number would have been enough.

4) Correct, but with emojis: This is a correct response as far as the location of the room is concerned. As before, there are two criteria that are met. The first one is the proper identification of the floor on which the target person is located. The second is the proper identification of the room in which the target person is located within that floor. However, there may be emojis interspersed throughout the conversation such as smileys. This was not requested of the LLM and morever it may not even be ideal for the purpose of a receptionist chatbot depending on the use case.

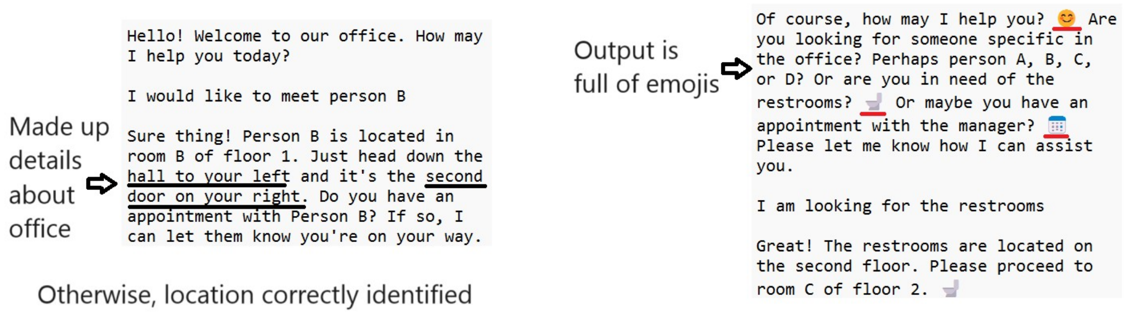


Figure 6: Left) An example where the answer is correct, but the layout of the office is assumed without being told by the LLM. Right) An example where the answer is correct, but is full of emojis not requested by the user

5) Half information: This sort of response does not give the complete answer. It does give a bit of the correct response, but not the full response. We need two criterial to be fulfilled if our response is to be considered useful for the user. The first one is the proper identification of the floor on which the target person is located. The second is the proper identification of the room in which the target person is located within that floor. With this kind of response, we can either be given just the floor number, or just the room number. In some cases, we can just be told that the target room is not on the current floor and thats it. This is not an overall useful response.

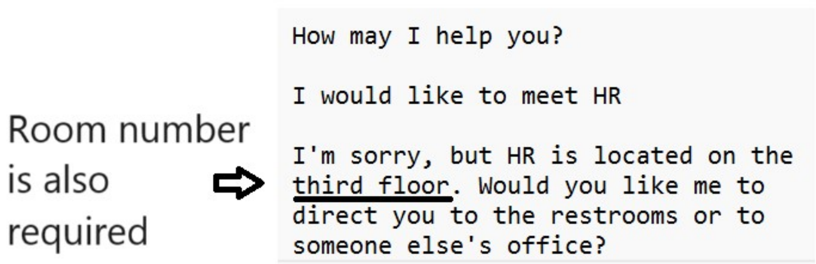


Figure 7: An example of only half information provided. Both room and floor number must be provided for a complete response.

6) Appointment issues: This sort of response has issues with determining whether the user should be asked if an appointment has been made or not. In our environment, the user should be asked if he has made an appointment before being told the location of the manager. If there is no appointment, then the location should not be disclosed. However, in some cases, the model asks the user regarding an appointment even though the manager is not being asked about. This is not the appropriate response since for anyone except the manager, the user should get the location without any further questions.

7) Incorrect response: This sort of response is incorrect. The location may not be accurate, or the location might not be given at all. A response is also considered inaccurate

if the location of the manager is given without confirming that an appointment has been made. This is not the ideal response for the user at all.

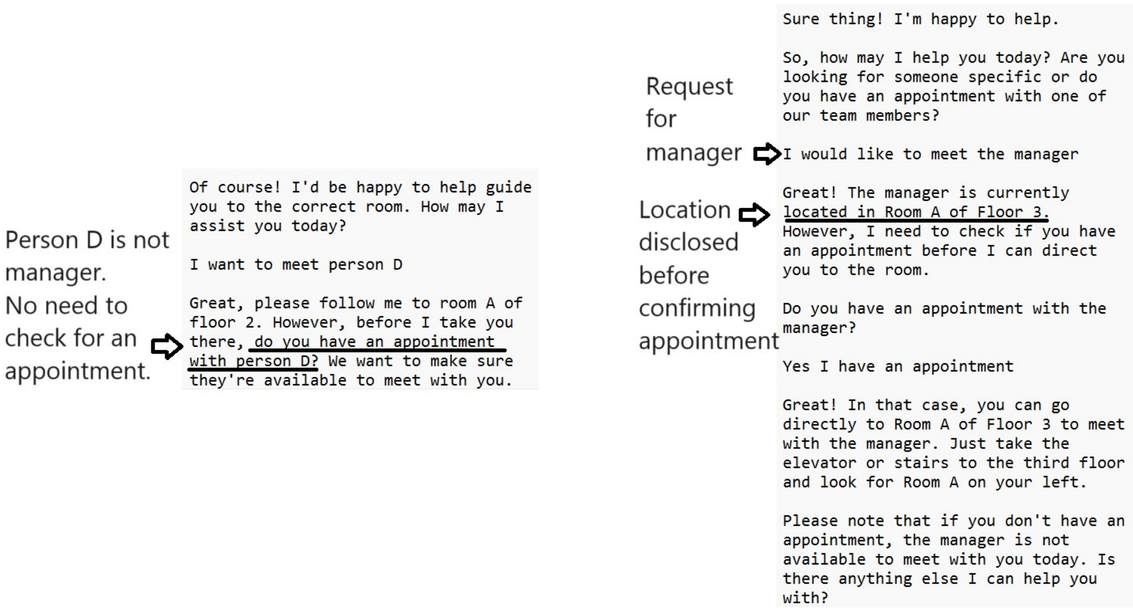


Figure 8: Left) An example of an issue with taking appointments. According to the prompt, the chatbot should only ask for an appointment if the manager is being inquired about. Right) An example of an incorrect response. The location of the manager is disclosed without confirming whether an appointment has been made or not.

Among all the answers, we also have a rare example of a situation where the chatbot assumes that we can be shown around by simply narrating it. This is counted as an example of half information provided as it does correctly identify the correct floor the target location is located on. The Vicuna 7b model gave this response.

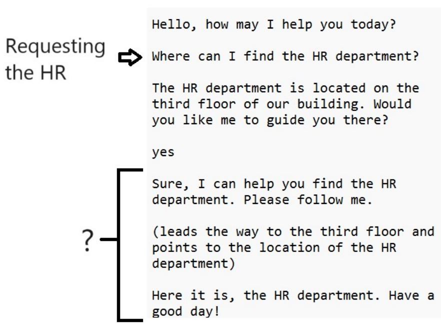


Figure 9: The chatbot assumes that we can be shown around by narration.

4 Implementation

A word cloud is a special type of image that is based on the amount of each word present in a text. It is basically made entirely of words, with the biggest word being the word most used in the corpus of texts. We can use the word cloud technique to visualize exactly what sort of words are most comm only used in the outputs presented by the LLM models.

This is an important step since such a visualization would allow us to guage if the chatbot really is using vocabulary that is appropriate for use in a receptionist setting. It may be possible that the model is making nonsense answers that do not help the user at all.

The way to do this is to make a seperate analysis for each of the LLM model that we are using. Since each model is different and likely has different training data, we cannot deny that the outputs to the same prompts would also likely be very different.

We can see an example of this in the wordcloud for the Vicuna 33b model. We can see the use of the word "elevator" here which was not prompted by the user as part of the environment being simulated. The model inferred that the user might reach the target floor by using an elevator by itself. Further prompting might be needed if the model is not to answer in such a way and stairs are the only way of reaching the location.

In the same way, when we look at the LLaMA 7b model wordcloud, we can see the use of the word "pauses". The LLaMA model has a different way of presenting outputs than the Vicuna model. It gives bits of text here and there that show "context". This can be something like a narration of what the receptionist is doing. As an example, the context may be "pauses and looks around", signalling that the receptionist is doing so while or before giving an answer. This may be considered noise or a useful part of the output depending on what the final use case of the model is.

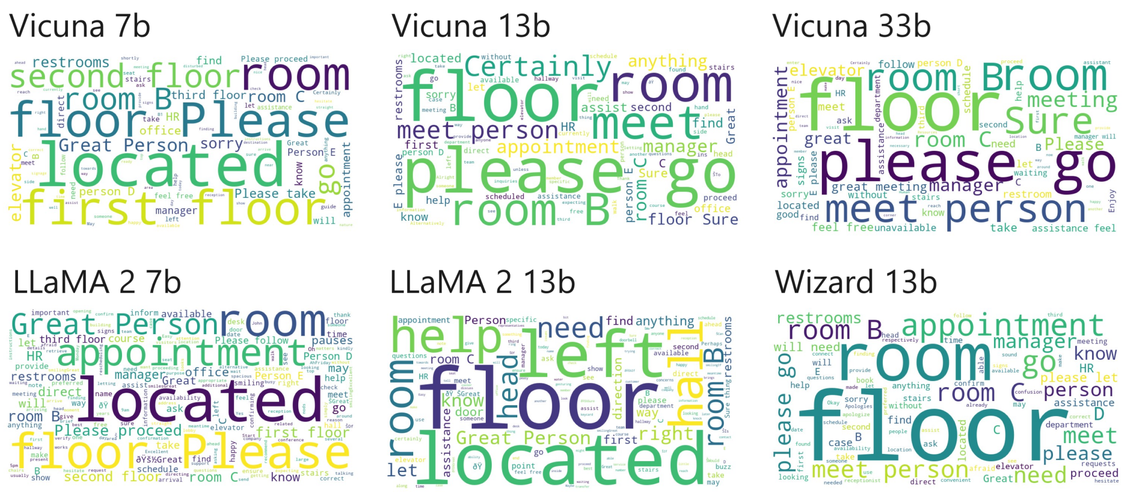
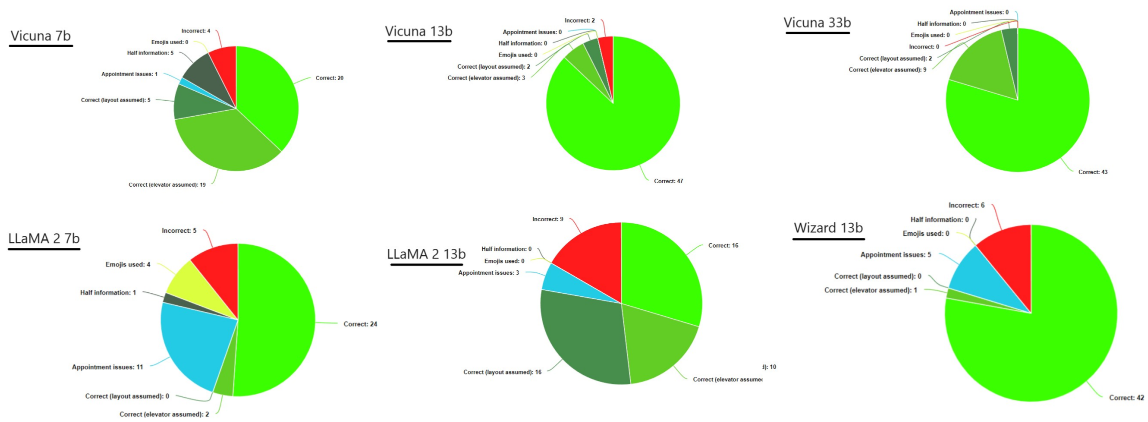


Figure 10: Wordcloud generated bu various LLM responses.

The outputs from more than 300 conversations were tabulated. We can come to sev-

eral possible conclusions that can be inferred from the results.
1) The number of correct responses increase as the size of the model increases. The Vicuna 7b model has only half the number of correct responses as the Vicuna 33b model. The number of incorrect responses from the Vicuna 7b model are also much higher than the other Vicuna models. In fact, the Vicuna 33b model does not have any incorrect responses.
2) Some models seem to be better suited for the receptionist chatbot role on first glance. The Vicuna and Wizard models give much more correct responses. On the other hand, the LLaMA models seem to give more responses with assumed elevator and assumed building layout segments in their answers. Also, the responses with emojis are almost always received from the LLaMA models alone. Vicna model doesnt appear to include emojis in its answers. The issue with asking for appointments for target locations other than the managers office also seems to be an issue mostly present in the LLaMA model as well
3) 13b models seem to have a good number of correct responses compared to the smaller ones. The level of improvement is big, with nearly a twice as many correct answers with the increase in parameters.



	Vicuna 7b	Vicuna 13b	Vicuna 33b	LlaMA 2 7b	LlaMA 2 13b	Wizard 13b
Correct	20	47	43	24	16	42
Correct (elevator assumed)	19	3	9	2	10	1
Correct (layout assumed)	5	2	2	0	16	0
Appointment issues	1	0	0	11	3	5
Half information	5	0	0	1	0	0
Emojis used	0	0	0	4	0	0
Incorrect	4	2	0	5	9	6

Figure 11: Responses tabulated in terms of accuracy.

Additionally, In terms of sentiment analysis, we find that all the LLM models give a positive sentiment value.

5 Evaluation

The results are varied. Some 13b models outperform others. In addition, atleast it can be observed that the performance within each model increases with general increase of parameters. Surprisingly, even this is not always true since for the LLaMA models, the 7b model had less incorrect answers than the 13b model. The various results can be tabulated and observed as follows:

	Vicuna 7b	Vicuna 13b	Vicuna 33b	LlaMA 2 7b	LlaMA 2 13b	Wizard 13b
Correct	20	47	43	24	16	42
Correct (elevator assumed)	19	3	9	2	10	1
Correct (layout assumed)	5	2	2	0	16	0
Appointment issues	1	0	0	11	3	5
Half information	5	0	0	1	0	0
Emojis used	0	0	0	4	0	0
Incorrect	4	2	0	5	9	6

Figure 12: Responses tabulated in terms of accuracy in a table format

5.1 Word Cloud

The word cloud demonstrates the relevancy of the answers provided. Almost all word-clouds generated focus on a few main keywords being the most used and this is ideal since it means that the different LLM models are at the very least all trying to answer exactly the question of where a particular person is in the office building. Words like "Room", "Located", "appointment", and "floor" are among the most common words in several of them. The one place where this changes is in the LLaMA models where the chatbor sometimes narrates what its doing by signalling that it "looks around" and "smiles". This is not observed in any of the other models. The LLaMA model is also the only model that gives outputs full of emojis, but this is not shown in the word cloud.

5.2 Sentiment analysis

The sentiment analysis is used to gauge if the chatbot has the proper attitude and does not give rude answers. We can see that all the models give a positive sentiment analysis result. The textblob library was used for this.

LLM Model	Sentiment
Vicuna 7b	Positive
Vicuna 13b	Positive
Vicuna 33b	Positive
LLaMA 2 7b	Positive
LLaMA 2 13b	Positive
Wizard 13b	Positive

Figure 13: Results of Sentiment analysis

5.3 Tabulation

After Tabulating the responses, it seems that the most effective LLM in the tests has been the Vicuna 13b model. It has the most correct answers (all models were given used to generate around 50 conversations to test on). There are other models with a high ratio of correct responses as well. The overall worse performing model has been the LLaMA

13b model. The two can be compared side by side in Figure 12:

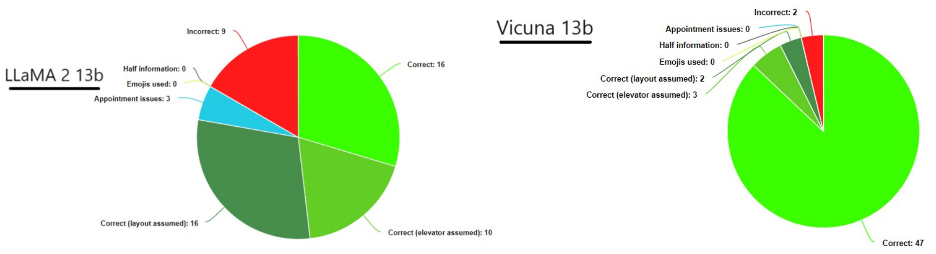


Figure 14: Worst performing model on the left, best performing on the right

6 Conclusion and Future Work

The research qestion was as follows: "How effective and comprehensive would currently existing AI and small scale LLM Models be in implementing a fully functional receptionist chatbot that needs to fulfil the tasks required to satisfy customer queries?"

It is observed that the existing LLM have many varying capabilities that can serve the needs of a receptionist quite well, however there is still a need for further improvement when it comes to generating the most accurate responses. Each model seems to generate a different sort of response and all of them can be further trained to generate responses different from the other. Further refinement of the existing models can improve the performance to the extent that the tasks performed have no issues at all.

A way of improving existing issues would be to continuously train the models on conversations and prompts that reinforce the layout and design of the office building in question so that the chatbot does not get confused and generate outputs which do not match with the actual layout.

LLM generated chatbot conversations can be fed once again into the LLM as further training data so that conversations only relevant to the building are fed into the LLM again and again until it properly matches each room to its correct location. This sort of work can also be extended to larger LLM models so that buildings of greater scale can be virtualized as well. An immediate further task that may be performed might be to add in a much bigger prompt that explains a much bigger office building and then test to see if the LLM responses are as good or not in that case.

To come up with future work and improvements, we can also take a look at what other research is looking into. Several people are looking into guidelines to use in regards to the future use of LLM in the future (Kim, Chua, Rickard and Lorenzo, 2023). We also see attempts at how fine tuning can be done using reinforcement learning from human feedback (Bill and Eriksson, 2023). There has also been a look into the emotional well being aspect of interacting with an LLM chatbot (FERRARA, 2022). There has also been looks into how the issues of privacy can be better tackled when it comes to chatbots and LLM (Ghayyur, Averitt, Lin, Wallace, Deshpande and Luthi, 2023). There can be discussions regarding ethics involved in the use of LLM models as well (Bang, Lee and Park, 2023) (Mvondo, Niu and Eivazinezhad, 2023). One improvement that can be made to the usage of any LLM model is using the right prompts (Zamfirescu-Pereira, Wong, Hartmann and Yang, 2023).

7 Appendix (a study of differences between a response from two different sized variants of the same model)

Prompting a request for the manager must be subject to asking the user if there is an appointment or not. The location is only shared if there is an appointment. Otherwise, the location is not provided. Here is an example of such an exchange generated by the Vicuna 13b model (the initial prompt is the same as mentioned above):

CHATBOT: How may I help you?
USER: I would like to meet the manager
CHATBOT: Sure, I can help you with that. Do you have an appointment with the manager?
USER: yes
CHATBOT: Great! You can go to room A of floor 3 for your appointment with the manager. Is there anything else I can assist you with?

There are many ways this sort of conversation can play out. In the conversation provided above, we can see that the result was successful and the location of the manager was correctly provided but was only given upon confirming that there was an appointment.

7.1 Differences in the answers prompted by each model

We can observe that the responses provided by the various models can vary. There are some similarities, but in large, the different models can greatly vary in the sort of responses that we can acheive. It should be taken in great consideration whether the model being used is actually inductive to the sort of responses that we actually require.

7.2 The typical responses of the Vicuna 7b LLM

Here is an example of a response given by the Vicuna LLM:

CHATBOT: Hello! How may I help you today?

USER: I would like to meet person B

CHATBOT: Great! Person B is located in room B of the first floor. Please go up the stairs and take a left at the top. Room B is straight ahead.
"

Here is another response:

CHATBOT: Hello, welcome to our office. How may I help you today?

USER: I want to meet person D

CHATBOT: I'm sorry, person D is not available on this floor. Can I direct you to room A of floor 2?

We can analyse some common patterns. Firstly, we can see that the initial greeting may vary. It is however almost always very brief and assumes that the user just walked into the office building at the same time. Details like the elevator and stairs are automatically inferred by the chatbot. Further training of the bot with further prompts and commands might direct the chatbot to only include such details if the building in question has such amenities at all. The response at the end contains the location of the room being inquired about.

7.3 The typical responses of the Vicuna 13b LLM

The Vicuna 13b model has certain changes as compared to the Vicuna 7b model. Here is an example of such a conversation:

CHATBOT: How may I help you?

USER: I want to meet person B

CHATBOT: Sure, to meet person B, please go to room B of floor 1.

This model is similar to the previous one, but the answers are more succint in this example. Here is an example of a slightly larger conversation:

CHATBOT: How may I help you?

USER: I would like to meet person E

CHATBOT: Sure, to meet person E, please go to room B of floor 2.
Thank you for your inquiry. To meet with person E, please go to room B of floor 2. Is there anything else I can assist you with?

USER: I would like to go to the restrooms

CHATBOT: Certainly, the restrooms are located in room C of floor 2. Please proceed to that room. Let me know if you need any further assistance.

Here we can again see that the directions to the location being inquired about are given at the end. The answers are simple and to the point.

7.4 The LLAMA 7b model

The LLAMA 7b model is prone to giving slightly larger responses. It also uses context a lot more. That is to say that in addition to the actual text responses to the question, the output also contains slight descriptions of the chatbots virtual behaviours and mannerisms such as smiling and looking around. Here is such a response:

CHATBOT: Of course, how may I help you? Do you need directions to a specific room or person in the office?

USER: I would like to meet HR

CHATBOT: Great! HR is located on the third floor. Would you like me to direct you to their office? pauses and looks around I'm afraid I can't just give you the address without confirming that you have an appointment. May I ask, do you have an appointment with HR? smiles

Note that the chatbot responses contain phrases like "pauses and looks around". These sorts of phrases are additional features of the outputs received from the LLAMA LLM model. Such responses are not part of the Vicuna model.

References

Arteaga, D., Arenas, J., Paz, F., Tupia, M. and Bruzza, M., 2019, June. Design of information system architecture for the recommendation of tourist sites in the city of Manta, Ecuador through a Chatbot. In 2019 14th Iberian conference on information systems and technologies (CISTI) (pp. 1-6). IEEE.

Bagchi, M., 2020. Conceptualising a Library Chatbot using Open Source Conversational Artificial Intelligence. DESIDOC Journal of Library Information Technology, 40(6).

Barbosa, A. and Godoy, A., 2021. Augmenting customer support with an NLP-based receptionist. arXiv preprint arXiv:2112.01959.

Bang, J., Lee, B.T. and Park, P., 2023, August. Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI. In 2023 International Conference on Platform Technology and Service (PlatCon) (pp. 109-113). IEEE.

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H. and Hemphill, L., 2023. A bibliometric review of large language models research from 2017 to 2023. arXiv preprint arXiv:2304.02020.

Bill, D. and Eriksson, T., 2023. Fine-tuning a LLM using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application.

FERRARA, A., 2022. Empowering emotional well-being through a LLM-based chatbot: a comparative study with the standard journaling technique.

Ghayyur, S., Averitt, J., Lin, E., Wallace, E., Deshpande, A. and Luthi, H., 2023. Panel: Privacy Challenges and Opportunities in LLM-Based Chatbot Applications.

Kim, J.K., Chua, M., Rickard, M. and Lorenzo, A., 2023. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology.

Lam, K.Y., Cheng, V.C. and Yeong, Z.K., 2023. Applying Large Language Models for Enhancing Contract Drafting.

Li, C.Y. and Zhang, J.T., 2023. Chatbots or me? Consumers' switching between human agents and conversational agents. Journal of Retailing and Consumer Services, 72, p.103264.

Lin, C.C., Huang, A.Y. and Yang, S.J., 2023. A review of ai-driven conversational chatbots implementation methodologies and challenges (1999–2022). Sustainability, 15(5), p.4012.

Liu, C., Jiang, J., Xiong, C., Yang, Y. and Ye, J., 2020, August. Towards building an intelligent chatbot for customer service: Learning to respond at the appropriate time. In Proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery Data Mining (pp. 3377-3385).

Lokman, A.S. and Ameedeen, M.A., 2019. Modern chatbot systems: A technical review. In Proceedings of the Future Technologies Conference (FTC) 2018: Volume 2 (pp. 1012-1023). Springer International Publishing.

Mvondo, G.F.N., Niu, B. and Eivazinezhad, S., 2023. Generative Conversational AI And Academic Integrity: A Mixed Method Investigation To Understand The Ethical Use of LLM Chatbots In Higher Education. Available at SSRN 4548263.

Pataranutaporn, P., Danry, V., Blanchard, L., Thakral, L., Ohsugi, N., Maes, P. and Sra, M., 2023, March. Living Memories: AI-Generated Characters as Digital Mementos. In Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 889-901).

Peng, B., Li, C., He, P., Galley, M. and Gao, J., 2023. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.

Suta, P., Lan, X., Wu, B., Mongkolnam, P. and Chan, J.H., 2020. An overview of machine learning in chatbots. International Journal of Mechanical Engineering and Robotics Research, 9(4), pp.502-510.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Wan, M. and Chen, X., 2018. Beyond" How may I help you?": Assisting Customer Service Agents with Proactive Responses. arXiv preprint arXiv:1811.10686.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Wong, A., Lacey, V., Gharpure, C., Hao, R., Venkatraman, P., Elidan, G., Engelberg, R., Hackmon, L., Rabin, R., Fink, M. and Yu, P., 2023. Reading Comprehension Assessment Using LLM-based Chatbot.

Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B. and Yang, Q., 2023, April. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-21).

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.