

1

# CSC413 Final Project Report

2

## 413 Teaching Assistant (TA) Chatbot

3

Emad Raheel Ahmad	UofT
Gaurav Singh Poona	UofT
Mohammed Hamlili	UofT
Nawal Kumar Mehta	UofT

### Abstract

4  
5  
6  
7  
8  
9  
10  
11  
12

The rapid advancements in AI have led to the development of numerous large language models (LLMs) such as OpenAI’s GPT-4, Google’s PaLM, and Falcon by Technology Innovation Institute. Meta AI’s recent introduction of the LLaMa 2 family further expands this landscape. This project aims to examine the performance of LLaMa 2 based chatbot in comparison to other leading LLMs. A key focus is exploring the feasibility of fine-tuning these models to compensate for their relatively smaller number of parameters. This research contributes to the discourse on LLMs, and also lays the groundwork for individuals to design and build their own chatbots easily.

# 1 Introduction

The principal objective of this research endeavor is to evaluate the feasibility of the LLaMa 2-Chat models as chatbots. Compared to two of the most widely regarded “best LLMs”, GPT-3.5 and GPT-4, the LLaMa 2-Chat models were found to be almost as accurate and powerful (Kadous, 2023), prompting further exploration. To substantiate this claim we plan to fine-tune the LLaMa 7b-Chat model on a chatbot dataset to improve its performance, and gauge if a well-designed dataset can compensate for the lower number of parameters in the LLaMa 2 models. We also compare the LLaMa 7b-Chat model to a model of similar capacity, namely the Falcon 7b model, which serves as the baseline model for our purposes.

Notably, the LLaMa 2 models are open source making them much more accessible than other LLMs and are also much smaller (the largest LLaMa 2 model has 70 billion parameters compared to an estimated 175 billion for GPT-3.5 and over 1 trillion for GPT-4). Thus, they require significantly less compute to train, making them much more cost effective than other available LLMs. If the LLaMa 2 models are able to keep up with their competitors in terms of performance, it opens the door for individuals wanting to build and fine-tune chatbots for their own specific purposes. One potential use for these models that we explore in TA chatbots for courses.

These models exhibit remarkable ability in learning intricate patterns in data and generating responses that resemble human communication. The surprisingly expressive power of the smaller LLaMa 7b-Chat model, which is based on an optimized transformer architecture, presents itself as the ideal candidate for this task. With only 7 billion parameters, it is the smallest of the LLaMa 2 family of models making it very easy to train, and its performance results will only be further improved upon by the larger LLaMa 2 models.

Large Language Models (LLMs) find extensive application in contemporary chatbots, particularly exemplified by the prevalence of OpenAI’s GPT-3 and GPT-4. In our increasingly technological landscape, chatbots play diverse roles, addressing a wide array of use cases. The integration of LLMs into chatbot frameworks stands as a well-established and proven solution, validated through extensive testing and real-world applications, and the LLaMa models could make this easier than ever.

## 2 Background and Related Work

### 2.1 Background

For the purposes of this paper, much additional background information beyond what we have learned in class is not required as our model architecture centers around concepts and techniques discussed in class. Namely, tokenization, RNNs, LSTM, encoders and decoders are used in preprocessing and in our model; which are all contained within our course content. Although, the actual model and baseline model we are using are not. That is the LLaMa7b and Falcon models, respectively. As well as some techniques used to load and train data in constrained environments.

#### 2.1.1 LLaMa7b:

LLaMa7b is a large open source language model developed by Meta AI, and is part of the LLaMA collection of foundational language models which ranges from 7 billion to 65 billion parameters. These models are trained on trillions of tokens from publicly available datasets with a neural network based around the transformer architecture, while also leveraging improvements proposed by other top NLP models. This includes pre-normalization from GPT-3, SwiGLU activation from PaLM, and rotary embeddings from GPTNeo. Ultimately, LLaMa-13b outperforms GPT-3 on most benchmarks and LLaMA-65b is competitive with the best NLP models like Chinchilla-70B and PaLM-540B. LLaMa-7b provides the same model architecture with fewer parameters, although this results in lower accuracy, it is also far faster and requires less compute resources. Thus making it the ideal version for the purposes of this project.

### 59 2.1.2 Falcon7b:

60 Falcon-7B is a powerful causal decoder-only model developed by the Technology Innovation Institute  
61 as part of the Falcon family of language models (also had 40b version). It is trained on a substantial  
62 1,500 billion token dataset; demonstrating superior performance to similar models such as MPT-7B.  
63 It is quite comparable to LLaMA7b in both size and function, therefore it will be used as our baseline  
64 model to evaluate our performance against.

### 65 2.1.3 Quantization:

66 Quantization is the process of reducing the precision of weights, biases and activations from higher-bit  
67 to lower-bit representation (e.g. 32-bit floating to 8-bit integers). This is done as it significantly  
68 decreases memory requirements and improves computational efficiency, which is ideal in resource  
69 constrained environments such as our Google colab notebook where we have limited RAM memory  
70 and slower GPUs. Therefore, we will be using the BitsAndBytesConfig when loading our model.

### 71 2.1.4 Lora:

72 Lora (Low Rank Adaption) is a technique used to accelerate fine-tuning of large models while also  
73 using less memory. This is done by using representing weight updates with two smaller matrices  
74 (update matrices), as this will reduce the total number of trainable parameters.

### 75 2.1.5 TRL:

76 Trl, built upon the transformers library by Hugging Face, is a library designed to train transformer-  
77 based language models; and supports most of decoder and encoder-decoder architectures. From  
78 this, we will be using the SFT (Supervised Fine-Tuning) Trainer, which allows easy fine-tuning of  
79 language model on custom datasets.

## 80 2.2 Related Work

81 Numerous studies have explored the application of neural networks and deep learning techniques in  
82 order to develop chatbots. Tsakiris, et al. (2022) performed a comprehensive analysis of several CNN  
83 model architectures as classifiers for the creation of chatbots; including AlexNet, LeNet5, ResNet  
84 and VGGNet. This was done by tokenizing raw input text, converting this text into vectors using  
85 “Doc2Vec” word embeddings, and then applying the model on those vectors to produce a prediction.  
86 These architectures were compared for their accuracy, f1 score, training time and execution time,  
87 highlighting significant differences in performance. Ultimately, LeNet5 proved to be the superior  
88 architecture with higher accuracy, faster training time and lower losses; despite having low accuracy  
89 on smaller datasets. In another study, Bhagwat (2018) conducted an in depth review of the several  
90 techniques used for NLP and chatbot development. Stating that RNNs (LSTMs and Seq2Seq models  
91 in particular) are desirable for chatbots as they have internal memory which allow them to deal with  
92 sequential data, which is needed to retain context in conversations. Bhagwat (2018) then proposes  
93 a hybrid model for chatbots, using a retrieval-based system, the context capabilities of LSTMs  
94 and the ease of use of a Seq2Seq model. This model is tested and compared against others, by  
95 preprocessing data, applying the model and using metrics such as bleu score and the turing test to  
96 measure performance. In the end, a hybrid model between a retrieval based and generation based  
97 system produces the best results. When compared to our architecture, we will be using similar  
98 preprocessing steps by tokenizing our inputs and converting text using word embeddings, although  
99 we will be using the LLaMA7b model architecture instead. LLaMA7b is an transformer based model  
100 that uses an attention mechanism; and often outperforms and is competitive against other leading  
101 NLP models, which is why we chose this over other models such as LeNet5.

## 102 3 Data

103 The foundation of our chatbot project is the robust dataset that supports the model's functionality  
104 and responsiveness. This section offers an extensive overview of the data gathered, in addition to  
105 the meticulous steps we took to ensure its cleanliness and compatibility with our deep learning  
106 architecture. We hope to convey the complexities and challenges of the task by providing insight into  
107 our data collection and preliminary processing procedures. Our description aims to project that we  
108 have thoroughly investigated the dataset's features and are well-equipped to identify and convey key  
109 information.

### 110 3.1 Dataset Description:

111 Our dataset is both comprehensive and purposefully diverse, containing a variety of prompts and  
112 responses based off of real life chatbot responses in several different languages, which functions  
113 collectively to improve the chat bot proficiency. The Hugging Face Datasets Hub provided us with  
114 our dataset, specifically the Guanaco LLaMa-2 1k dataset. This dataset takes 1000 training datapoints  
115 from the openassistant guanaco dataset and applies a transformation process upon it, such that the  
116 resulting datapoints are in the correct format accepted by the LLaMa model (Hugging Face, 2022).  
117 Therefore, the resulting dataset contains 1000 datapoints where each datapoint simply contains  
118 a singular field. This field contains the example question as well as its corresponding response;  
119 formatted as follows: "<s> [INST] prompt [/INST] response <s>" as this is the format accepted by  
120 LLaMa.

### 121 3.2 Data Preprocessing:

122 Our rigorous data preprocessing illustrates our devotion to data integrity and effectiveness. This  
123 comprised of extensive data cleaning and formatting to eliminate inconsistencies and ensure natural  
124 language understanding, and corresponding the data with the architecture for the model. We meticu-  
125 lously transformed the data to align its structure with the architecture that comprises the LLaMa-2  
126 model. This entailed removing all non-English datapoints, as for the purposes of our report we are  
127 only operating with the English language. Furthermore, we were required to format any dataset  
128 that we attempted to train with, such that it matched the desired format of LLaMa-2. This meant  
129 going through each data point and retrieving the prompt with its corresponding response, and then  
130 reformatting it into one field (if it wasn't already) in the format described previously. This means  
131 wrapping the prompt in "[INST]" and "[/INST]" tags and concatenating it with the response, placing  
132 the entire text between "<s>" and "</s>" tags.

### 133 3.3 Exploratory Figures:

```
<s>[INST] I want to start doing astrophotography as a hobby, any suggestions what could i do? [/INST] Getting started in  
astrophotography can seem daunting, but with some patience and practice, you can become a master of the craft. To begin, you'll need_  
  
<s>[INST] Método del Perceptrón biclásico: definición y variantes del método. ¿cómo se aplicaría al caso multiclásico?, ¿se podría  
utilizar con patrones que sólo son cuadráticamente separables? [/INST] El método del Perceptrón biclásico es un algoritmo de_  
  
<s>[INST] Schreibe einen kurze und präzise Konstruktionsbeschreibung zu einem Dreieck ABC mit c=6 cm, h_c=5 cm und γ=40°. γ ist  
hierbei der von Seite c gegenüberliegende Winkel. [/INST] Dreieck ABC ist ein rechtwinkliges Dreieck mit der Seitenlänge c=6 cm als_  
  
<s>[INST] Can you give me an example of a python script that opens an api point and serves a string? [/INST] Sure! Here's an example  
Python script that uses the Flask web framework to create a simple API endpoint that serves a string: ``` from flask import Flask ap_  
  
<s>[INST] How can I learn to optimize my webpage for search engines? [/INST] Learning to optimize your webpage for search engines  
involves several steps, including: 1. Understanding the basics of how search engines work: Learn how search engines crawl and index_
```

Figure 1: Excerpt from dataset

134 This is a brief excerpt from the dataset on which we fine-tuned the LLaMa 7b-Chat model. It shows  
135 the different samples with the prompt encased within the [INST] tags and the response after the  
136 closing [/INST] tag.

## 137 4 Model Architecture

138 We propose employing the Llama7b neural network architecture to develop the Chatbot, which is  
139 renowned for its capacity to execute natural language processing tasks. The Llama7b model both  
140 employs a transformer-based model architecture and comprises a deep recurrent neural network  
141 (RNN) alongside long short-term memory (LSTM) units which were chosen strategically for their  
142 ability to capture sequential dependencies within conversational inputs. (Dong et al., 2023)

### 143 4.1 Key Components:

- 144 • **Embedding Layer:** This component of the model utilizes word embeddings to repre-  
145 sent input text within a continuous vector space, which improves the model’s semantic  
146 comprehension.
- 147 • **Recurrent Layers:** The core is comprised of stacked LSTM layers that aid in the capture of  
148 sequential patterns and contextual dependencies within user interactions.
- 149 • **Attention Mechanism:** Using an attention mechanism to dynamically weigh various  
150 elements of the input, the model’s focus on salient information can be refined.
- 151 • **Output Layer:** The final layer generates probability distributions across the vocabulary,  
152 facilitating coherent and contextually appropriate textual responses to be generated.

### 153 4.2 Hyperparameters and Training Settings:

154 We used the Supervised Fine-Tuning Trainer class to run the training of the model on the chosen  
155 dataset.

156 Given the complexity of the model with 7 billion parameters, each epoch would take approximately  
157 25 minutes to run and would achieve remarkable results.

158 There was a number of hyperparameters to consider (such as batch size, number of epochs, learning  
159 rate, weight decay, etc) when training the model, and these were passed as arguments to the model’s  
160 training wrapper.

161 Thus, we were limited for the fine-tuning of the hyperparameters because a grid search over a range  
162 of various values on a single T4 GPU would take hours, if not more. We decided to use random  
163 search instead, where we adjusted the value of a single parameter each time and observed the new  
164 performance (some example attempts were included in our code). We noticed that the performance  
165 was similar for the different models, and the best model ran with a training loss of 1.529, where we  
166 multiplied the learning by 3.

### 167 4.3 Rationale for Choosing Llama7b:

168 The Llama7b architecture coincides with the primary objectives for our endeavour, which are  
169 demonstrating success in natural language understanding tasks. The model’s ability to engage in  
170 meaningful and contextually relevant conversations is strengthened through its recurrent structure  
171 and attention mechanism.

## 172 5 Model Architecture Figure

173 We produced a concise yet informative figure (Figure 1) that illustrates our proposed Llama7b model  
174 architecture, which elucidates the key components and information flow within the neural network.

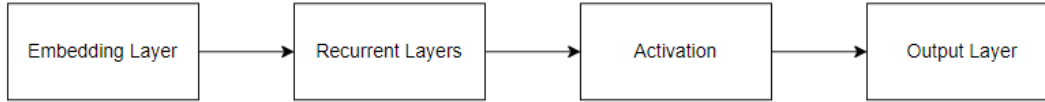


Figure 2: Model architecture

## 6 Results

Given that our model is a transformer's architecture that generates text, we decided to evaluate the model using qualitative assessments against [1] model answers on prompts the model has never seen before, [2] answers on similar prompts by a different model and [3] answers on similar prompts against GPT-3.5.

### 6.1 Evaluation against Test Set

We used the test subset of the "mlabonne/llama-2-7b-guanaco" dataset to evaluate the model. Since it has a lot prompts in multiple languages, we decided to qualitatively assess our model on 5 english prompts. We sampled these, and ran the model on each prompts and then compared it to the answer that was already provided in the testing dataset. This can be found in our output in the "ChatbotFinalProject.ipynb" notebook.

The first prompt proved that the model can provide similar questions than expected to general questions about modern tools like ChatGPT. The third prompt proved that the model can also generate correct code for a specific algorithm request (Bubble Sort, in this case) on a specific language (C, in this case). The other prompts also showed that the generative capabilities of the model are reasonable and provide coherent and grammatically correct text, all while responding to specific prompts.

### 6.2 Evaluation against Falcon 7b model

The responses of the Falcon 7b model and our model to the first prompt ("What is a large language model?") are shown in the figures below.

1. A large language model is a type of artificial intelligence (AI) that is trained on a large dataset of text to generate human-like language. everybody has their own definition of what a large language model is, but generally, it is a model that has been trained on a dataset of at least 100 million words.

Large language models are trained on large datasets of text, such as books, articles, and websites. They are designed to learn the patterns and structures of language, and to generate text that is similar to the training data.

Large language models are used in a variety of applications, such as chatbots, language translation, and content generation. They are also used in research to study the structure and evolution of language.

Some of the most popular large language models include:

1. BERT (Bidirectional Encoder Representations from Transformers)

Figure 3: Our model's response

1. A large language model is a type of machine learning algorithm that uses a deep neural network, typically a recurrent neural network or a convolutional neural network (CNN), to learn the mapping between two or more languages by extracting the features of one language (usually text) and applying it to another language. It can also be used to perform natural language understanding and processing tasks such as sentiment analysis, text classification, and translation.

Figure 4: Faclon 7b response

194 The Falcon 7b model's response was not as robust as our model's. It failed to mention where the  
195 word "large" comes from in large language models, and also was only able to provide simpler and  
196 perhaps outdated applications. This trend continued in the rest of the prompts as well, with Falcon 7b  
197 often not providing sufficient detail.

### 198 6.3 Evaluation against GPT-3.5 and GPT-4

199 We also evaluated our fine-tuned model against the most widely known LLM based chatbot models  
200 today, GPT-3.5 Turbo and GPT-4. The responses of GPT-3.5 Turbo and GPT-4 to the first prompt  
201 ("What is a large language model?") are shown in the figures below. Since our model output was  
limited to a sequence length of 200, we also set this restriction on GPT's outputs.

1. A large language model is a sophisticated artificial intelligence system designed to understand and generate human-like text. It is trained on vast amounts of diverse language data to learn patterns, context, and linguistic structures. These models, such as GPT-3.5, utilize deep neural networks with millions or billions of parameters, enabling them to process and generate coherent and contextually relevant text. They excel in tasks like natural language understanding, translation, and content generation, making them valuable tools for various applications, from chatbots and virtual assistants to content creation and language-related tasks. The model's ability to comprehend and produce nuanced text stems from its extensive training on diverse language samples, enabling it to exhibit human-like language capabilities.

Figure 5: GPT-3.5 Turbo response

202 Overall, GPT-3.5's responses were similar to that of our fine-tune LLaMa 7b-Chat model. For  
203 example, in the first prompt, both responses explained the general idea behind LLMs, the kind of  
204 data they are trained on, and their applications. Both chatbots also highlighted the ability of LLMs to  
205 generate human-like responses.  
206

1. A large language model (LLM) like ChatGPT is an advanced type of artificial intelligence that understands and generates human-like text. It's built using a technique called deep learning, specifically a form of neural networks called transformers. The "large" part refers to its massive scale, as it's trained on a vast amount of text data from books, articles, websites, and other sources. This training enables it to grasp a wide range of topics, understand context, and generate coherent and contextually appropriate responses. Its applications range from answering questions and writing essays to more complex tasks like translation, summarization, and even creative writing. However, it's important to remember that while LLMs are powerful, they rely on the data they've been trained on and can sometimes generate incorrect or biased information.

Figure 6: GPT-4 response

207 The responses of the GPT-4 model were more coherent and informative than our model's. In the first  
208 prompt, like GPT-3.5 and our model, GPT-4 also talked about idea of LLMs, how they are trained and  
209 their uses. However, GPT-4 also mentions that LLMs can be biased and incorrect at times, which both

210 GPT-3.5 and our model fail to mention. This trend continued in the other prompts as well. GPT-4  
211 was able to consistently provide just a little bit more relevant information beyond what was asked.

## 212 7 Discussion

213 In this section, we go over the performance given by our chatbot, which employs the LLaMa7b neural  
214 network architecture. The evaluation entails comparing our model's performance to a baseline and  
215 reflecting on its behaviour across various aspects.

### 216 7.1 Model Performance Assessment

#### 217 7.1.1 Baseline Comparison

218 The LLaMa7b model, an advanced language model utilizing a transformer-based architecture, is  
219 employed by our chatbot. The baseline model, LLaMa7b, offers an established foundation for natural  
220 language understanding tasks. We chose this model due to its recurring structure and attention  
221 mechanism, both of which coincide with our primary goals of illustrating success in contextually  
222 relevant conversations. The LLaMa7b outperforms simpler models like Falcon in terms of nuanced  
223 and effective conversational interactions.

#### 224 7.1.2 Model Behaviour Reflections

225 The LLaMa7b model conveys recurring trends and contextual dependencies within user interactions  
226 sufficiently. The integration of an embedding layer improves semantic comprehension by enabling  
227 the model to represent input text in a continuous vector space. The recurrent layers, composed of  
228 arranged LSTM units, help the model handle sequential data effectively.

### 229 7.2 Model Refinement and Future Directions

230 Future efforts will be centred on refining the model architecture in light of the observed strengths  
231 and limitations. Experimenting with various LLaMa7b models while delving into other models, such  
232 as RAG, can provide suggestions for potential improvements. Furthermore, incorporating real-time  
233 information retrieval mechanisms may address the limitations of static, pre-processed data.

## 234 8 Limitations

235 While the chatbot possesses promising capabilities, it is crucial to recognize certain limitations which  
236 could influence its performance:

### 237 8.1 Contextual Understanding

- 238 • **Lack of Real-Time Information:** Given that the chatbot is dependent on pre-processed  
239 data, responses may not reflect the most up-to-date information.
- 240 • **Contextual Nuance Sensitivity:** Despite advanced capabilities, the chatbot might have  
241 difficulty with nuanced or context-specific queries.

### 242 8.2 Dataset Restrictions

- 243 • **Data Bias:** The quality and diversity of the training dataset influence the chatbot's effective-  
244 ness, hindering its ability to respond to an extensive variety of queries.
- 245 • **Limited Domain Coverage:** Given that the chatbot's wisdom is confined to specific sources,  
246 its applicability beyond the defined domain could be limited.



### 247 8.3 Interactive Restrictions

- 248 • **Clarifications-based Responses:** Slang, or specialized terms could leave a bearing regarding  
249 the chatbot’s performance.

### 250 8.4 Limited Compute Resources

- 251 • **Working in Google Colab:** Free versions of Google Colab has extremely limited RAM  
252 memory as well as slower GPUs, which hinders training significantly as it prevents us from  
253 training on very large datasets, in higher precision and with greater overall accuracy. This  
254 ultimately affects our model performance all around, resulting in worse responses all around.

255 Acknowledging these constraints is essential for managing user expectations, and subsequent en-  
256 hancements could tackle these issues to make the chatbot more versatile and responsive.

## 257 9 Ethical Considerations

258 It is essential to address any potential ethical concerns associated with the utilization of our chat bot.  
259 It is entrusted upon us to engage in thoughtful deliberation regarding issues of privacy, fairness, and  
260 accountability in the context of machine learning models employed in different settings, including an  
261 educational context and general human-computer interactions. As we strive to enable the chat bot to  
262 engage in meaningful conversations regarding a variety of topics, as well as high level concepts such  
263 as machine learning, it becomes our responsibility to remain vigilant regarding protecting the user’s  
264 personal data and fostering respectful and inclusive interactions.

265 In this regard, we are devoted to preventing and addressing to any potential chat bot misuse. Spam-  
266 ming, harassment, and other forms of inappropriate conduct should not be tolerated in the environment  
267 we seek to create. To adhere to the values of transparency and user consent, we intend to implement  
268 mechanisms which ensure the ethical use of our model, guided by established principles and moral  
269 standards that prioritize our users’ well-being and educational experience.

270 In an educational context, it is critical to emphasize that any attempt utilize the chat bot into  
271 granting answers to assignment or test questions is strictly prohibited. Such behaviours not only  
272 undermine the integrity of the educational process but also breach academic integrity standards. We  
273 are committed to proactively discouraging and addressing any such behaviour in order to uphold a fair  
274 and academically rigorous environment. Finally, our project revolves around a strong commitment to  
275 ethical considerations and the promotion of academic integrity in all interactions between students  
276 and our chat bot.

## 277 10 Conclusion

278 In the end, our chatbot—which is built on the reliable LLaMa-2-7b language model—shows profi-  
279 ciency in conversing on broad and intricate subjects like machine learning. The chatbot uses a variety  
280 of statistics to fine-tune its responses to human inquiries and contextual nuances.

281 Our results emphasize the chatbot’s capability to generate coherent responses while also highlighting  
282 its adaptability to changes in machine learning research. Although there are limitations in handling  
283 nuanced queries and real-time information, our chatbot makes a big impact to the educational  
284 technology landscape by demonstrating how one can utilize LLMs for this purpose.

285 The project’s key finding is the value of using Large Language Models to create specialised teaching  
286 resources that give students an engaging environment in which to engage with material related to  
287 machine learning. We were able to baseline a specific open-source model and fine tune it against  
288 a general dataset, and compare it against other models. We anticipate further advancements when  
289 employing language models for chatbots as we refine the chatbot’s capabilities. Essentially, this  
290 chatbot symbolises the revolutionary potential of massive language models in predicting the direction  
291 of educational technology.

## 292 References

- 293 [1] Bhagwat, V. A. (2018). Deep learning for chatbots.
- 294 [2] Dong, Y., Lu, W. J., Zheng, Y., Wu, H., Zhao, D., Tan, J., ... Cheng, W. (2023). Puma: Secure inference of  
295 llama-7b in five minutes. *arXiv preprint arXiv:2307.12533*.
- 296 [3] Hugging Face. (Year). Guanaco LLaMa-2 1k dataset. Hugging Face Datasets.  
297 <https://huggingface.co/datasets/mlabonne/guanaco-llama2-1k>
- 298 [4] Kadous, W. (2023, August 23). Llama 2 vs. GPT-4: Nearly as accurate and 30x cheaper.  
299 Anyscale. [https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-](https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-is-30x-cheaper)  
300 [is-30x-cheaper](https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-is-30x-cheaper)
- 301 [5] Riedel, S., Piktus, A., Lewis, P., Kiela, D. (2020). Retrieval augmented generation: Streamlining the creation  
302 of Intelligent Natural Language Processing Models. AI at Meta. [https://ai.meta.com/blog/retrieval-augmented-](https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/)  
303 [generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/](https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/)
- 304 [6] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... Lample, G. (2023). Llama:  
305 Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- 306 [7] Tsakiris, G., Papadopoulos, C., Patrikalos, G., Kollias, K. F., Asimopoulos, N., Fragulis, G. F. (2022). The  
307 development of a chatbot using Convolutional Neural Networks. In *SHS Web of Conferences* (Vol. 139, p.  
308 03009). EDP Sciences.