Exploring the Behavior and Performance of Large Language Models: Can LLMs Infer Answers to Questions Involving Restricted Information?

Ángel Cadena-Bautista , Francisco López-Ponce , Sergio Ojeda-Trueba , Gerardo Sierra * , Gemma Bel-Enguix

Posted Date: 7 November 2024

doi: 10.20944/preprints202411.0502.v1

Keywords: RAG; Large Language Models; Information Retrieval; Bible corpus

 doi:10.20944/preprints202411.0502.v1

Article

Exploring the Behavior and Performance of Large Language Models: Can LLMs Infer Answers to Questions Involving Restricted Information?

Ángel Cadena Bautista [1,†], Francisco López Ponce [1,†], Sergio Ojeda Trueba [1,†], Gerardo Sierra [1,*,†] and Gemma Bel-Enguix [1,†]

Instituto de Ingeniería, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico; angelcaden@hotmail.com (A.C.B.); francisco.lopez.ponce@ciencias.unam.mx (F.L.P.); sojedat@iingen.unam.mx (S.O.T.); gbele@iingen.unam.mx.com (G.B.-E.)

* Correspondence: gsierra@iingen.unam.mx

† These authors contributed equally to this work.

Abstract: In this paper various LLMs are tested in a specific domain using a Retrieval-Augmented Generation (RAG) system. The study focuses on the performance and behaviour of the models and was conducted in Spanish. A questionnaire based on The Bible, which consist of questions that vary in complexity of reasoning, was created in order to evaluate the reasoning capabilities of each model. The RAG system matches a question with the most similar passage from The Bible and feeds the pair to each LLM. The evaluation aims to determine whether each model can reason solely with the provided information or if it disregards the instructions given and makes use on its pretrained knowledge.

Keywords: RAG; large language models; information retrieval; Bible corpus

1. Introduction

Large Language Models (LLMs) have transformed information access and processing, particularly in question-answering systems. These models understand and respond to queries in natural language, facilitating intuitive and accessible interfaces for users. Their ability to analyze context and produce relevant and coherent responses is crucial in areas such as legal [1], medical [2], and financial sectors [3].

These models, trained on vast datasets, posses the capability to answer specific queries by extracting and synthesizing knowledge from their internal databases. However, a significant challenge arises when the required information is restricted, meaning it contains specific details that might not be in their training database. For instance, in the case of technical manuals, such as construction regulations that feature different stipulations from one country to another, or company policies and procedures that may substantially differ from general regulations.

The central issue addressed in this paper is the evaluation of an LLM's behavior when asked to reason and answer based solely on a restricted piece of information. This information ideally is part of the model's training data, but is modified and questioned in a way that forces the LLM to focus on particular details that involve reasoning and precise answer identification, rather than a generic response.This distinction is crucial because it forces LLMs to adapt and combine training information with newly unseen content in a consistent manner.

In virtue of this, the challenge lies in these models' ability to interpret and apply their knowledge to cases that, while superficially similar to previously seen situations, differ in critical aspects that may affect the accuracy of the responses generated. For example, in the legal field, an LLM might be trained on the jurisprudence of a specific country but could be asked to apply those principles to a case in a different legal context, which, although similar, has unique regulations and precedents. The ability to adjust to these subtle yet fundamental differences is what we're going to observe, in order to optimize the effectiveness of LLMs in practical applications.

 doi:10.20944/preprints202411.0502.v1

This objective allows us to test information retrieval systems, meaning that although the main results will focus on the questionnaire and each language model's performance, various technical results will be reported since they give insight into the ins and outs of information processing with LLMs.

To conduct an effective evaluation aligned with our objectives, a set of questions was developed around a widely known domain, namely The Bible, with answers that must be derived solely from a specific set of restricted verses. The biblical domain was selected as the central theme due to its immense cultural impact and the high quantity of information available online, meaning that any respectable LLM must be able to answer basic questions about it. Nonetheless, this domain is vast enough to be able to generate very precise questions that can't be answered solely with general knowledge. The questions themselves were created so that they have to be answered based on very particular verses, and inferences based on them, more on that later.

This study not only deepens our understanding of the current capabilities of LLMs, but also raises crucial questions about how these models might be designed or modified to effectively handle selective information. Such considerations have significant implications for the design of artificial intelligence systems that must operate in restricted environments and handle sensitive information. This is particularly important given the increasing amounts of industry-level applications and personal assistants that use LLMs.

The work is structured as follows. Section 1 introduces the background and motivation for this study. Section 2 reviews related work in question answering and retrieval-augmented generation (RAG) systems. Section 3 provides a detailed description of the models employed in this research. Section 4 outlines the characteristics of the corpus and the questionnaire used. Section 5 details the methodology used throughout the study. Section 6 presents the results, followed by a discussion of these findings in section 7. Finally, section 8 concludes with a summary of the key conclusions.

2. Related Work

The resources required to train an LLM prevent these models from easily expanding their memory, leading to what are called "hallucinations" [4] when generating unknown content. Hallucinations are factually incorrect answers to a particular query disguised with a correct use of grammar. Hallucinations generate problems when working with retrieval of information in specialized texts [5].

To address this issue, hybrid systems have emerged aiming to find ways for these models to review previously unseen information and expand their knowledge before generating answers. Some examples include REALM [6], ORQA [7], and RAG [8].

Retrieval Augmented Generation (RAG) is a widely used LLM enhancement technique that relies on filtered and reviewed information to modify and optimize an LLM's answer to a given question. RAG is a method of expanding the model's knowledge by analyzing a given query, retrieving related information from a particular database, and feeding the LLM with the retrieved text before the generation of the answer. RAG systems have shown that they can reduce factual errors within a model's response to a question [9]. This demonstrates that LLMs are capable of integrating retrieved information into their responses and modify them accordingly.

Thorough research has been carried out regarding RAG. Gao [10] classifies these systems in three categories: *Naive RAG*, *Advanced RAG*, *and Modular RAG*, based on how complex each building block of the system is. Naive RAG corresponds to unified systems that, using each LLM's framework, carry out a vanilla indexing, embedding, retrieval, and generation. An example of this system can be found in [5]. Advanced RAG builds upon the Naive architecture but alters the retrieval process, adding information such as tags and text-relevant metadata as a pre-retrieval step, as well as a post-retrieval adjustment that filters and rearranges the retrieved information in order to avoid information overloading for the LLM. These previous classifications preserve a linearity of the information flow with the system, allowing a precise analysis of each building block. The RAG system discussed in this paper corresponds to an Advanced RAG system.

 doi:10.20944/preprints202411.0502.v1

The final category is Modular RAG, which utilizes various auxiliary modules to modify the system's functionality in multiple ways. Certain modules can rephrase the input prompt to generate different perspectives of the same question before providing it to the LLM. Some extend retrieval to various data sources like search engines, traditional databases, and even graph-based sources. Others adjust new responses based on embedding similarities between previous answers and real-world documents. Such a model can be seen in [11], where an SQL database is implemented in order to generate queries for a RAG system.

Regardless of the particular RAG variation, this methodology is widely used in order to improve an LLM's performance in Question Answering systems over various different fields of knowledge. Implementations and benchmarks of these types of systems tend to be focused on freely accessible information like medicine or law corpora. In [12] a medicine field RAG model is presented with a state of the art evaluation comprised of over 7000 questions from 5 medicine related datasets, the RAG system improves Accuracy in this test up to 18% in LLMs. In [13] an *Advanced RAG* implementation is carried out over a law questionnaire, focusing on various information retrieval methodologies in the retrieval aspect of the model and obtaining a 90% similarity when comparing generated responses and a human created gold standard, in over 2000 questions.

Regarding QA systems utilizing The Bible as a source of information, limited research has been conducted, with the most notable being [14]. In this paper, a Bible-related QA dataset was created, and an extractive QA model based on recurrent and convolutional neural networks, along with domain adaptation, was implemented. The model was tasked with answering Bible-related questions by analyzing various Bible chapters in order to select the one containing the answer to the question. This paper extends the aforementioned project using contemporary resources, opting for a RAG system instead of domain adaptation and employing LLMs instead of recurrent or convolutional networks. Due to the increasing impact of LLMs, this paper shifts its focus to behavior analysis rather than solely conducting a Question Answering evaluation. Nonetheless, several elements are shared between the studies, such as the creation of a dataset consisting of questions and chapters from which the correct answer can be extracted.

3. Models

In this section, we describe the models used for our experiment. This RAG implementation is divided in the two standard sections: retrieval and generation. For the retrieval we apply the BGE-M3 embedding model [15] to our database and the questions, based on these embeddings the semantic search is carried out. For the second part, answer generation, we use three state of the art LLMs: Llama 2, in its fine-tuned 13B chat version, [16], GPT, in its 3.5 version [17] and PaLM. [18].

3.1. BGE-M3 Embeddings

The BGE-M3 Embedding (BGE-M3) model [15] is a multi-lingual, multi-functionality and multi-granularity model, that can support more than 100 languages. It unifies common retrieval functionalities of text embedding models, making it able to generate various modified embeddings from the same text, in particular three different categories: dense, lexical, multi-vector.

There are certain benefits of using BGE-M3 over classic SentenceTransformer-based models. BGE-M3's multi-granularity enables it to work with inputs of varying length, having a max length of 8,192 tokens. This is particularly useful considering the length of the textual information with which we will be working with (refer to section 4). As much as we're not focusing on LLMs' multilingual capabilities, the version of The Bible that we're using as well as the questionnaire are in Spanish, hence having language independent embeddings is vital.

Finally we want to test the impact of the different categories of generated embeddings in a RAG system. BGE-M3's dense vectors are fairly similar to most encoder based methods: a text t is initially embedded using an encoder model, yet t's corresponding embedding is the normalized hidden state representation of the CLS token. Similarity between these vectors is standard inner product. BGE-M3's

 doi:10.20944/preprints202411.0502.v1

new embeddings are lexical embeddings and multi-vector embeddings. Lexical embeddings generate a weight for each term in the sentence using a RELU activation over a hidden state of the neural network, where similarity between these vectors is obtained between reoccurring terms in both vectors. Multi-vector, similarly to dense embeddings, consists of a normalized encoder based embedding. The main differences is the use of a projection matrix to adjust the embedding, before obtaining a point wise score from the dot product of said embeddings.

As they explained in their paper, the model was trained over three different sources; unsupervised data from unlabeled corpora, fine-tuning data from labeled multilingual corpora, and synthetic LLM generated data.

The training objectives of the three different embedding vectors can be mutually conflicting and to facilitate the optimization the training process is on top of a self-knowledge distillation. Based on the principle of ensemble learning, the heterogeneous predictors can be combined as a stronger one.

The model was evaluated on three tasks: multi-lingual retrieval, cross-lingual retrieval, and multi-lingual long document retrieval.

3.2. Llama 2

Llama 2 [16] is an updated version of Llama 1 [19], developed by Meta Research and open-source. Llama 2 was trained on a new mix of publicly available data. Compared to Llama 1, the the size of the pretraining corpus in Llama 2 was increased by 40%, doubled the context length, and adopted grouped-query attention. Llama 2 was released in three variants with 7B, 13B, and 70B parameters. For this paper, we use Llama 2 13b-Chat, the fine-tuned version of Llama 2, which is optimized for dialogue.

We opted for the chat optimized version of the model since various training modifications were carried out in order to perfect the model's response capabilities. Llama 2 Chat is a fine-tuned version of Llama 2 improved for dialogue cases that aims to increase helpfulness and safe answers. The technical modifications are supervised fine-tuning trough the compilation of self annotated data. Subsequently, the model is iteratively refined using Reinforcement Learning with Human Feedback (RLHF) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO).

Llama 2 new data corpus was cleaned to avoid including data from Meta's products or services. Data from certain sites that are known to contain a high volume of personal information were removed. Additionally, the most factual sources were up-sampled to increase knowledge and damped hallucinations. This fact is to be tested in the various experiments presented in this paper.

At the time of release, Llama 2 surpassed various open-source models and competed with most closed-source models in benchmarks for World Knowledge, Reading Comprehension, as well as Mathematics evaluations.

3.3. PaLM

Before Google's introduction of the Gemma and Gemini series of LLMs, PaLM [18] was the company response to state-of-the-art LLMs. PaLM is a dense decoder-only model with 540 billion parameters trained on 780 billion tokens. This model was efficiently trained, particularly in terms of scaling, using pathways that enable training a single model across multiple chips in a highly efficient manner.

As per most pretrained models, improvements can be seen when scaling model sizes. The larger models have better performance, and PaLM is no exception given the amount of training data used. The best resulting version of PaLM (when evaluated in several natural language tasks) is that which has the largest amount of parameters. This model is evaluated in multilingual benchmarks for machine translation, summarizing and question answering. Similarly bias and toxicity are evaluated analyzing whether PaLM is prone to affirm stereotypes on gender, race, occupation or religion.

 doi:10.20944/preprints202411.0502.v1

3.4. GPT

GPT, in its 3rd version presented in [17], is an auto-regressive language model trained with in-context learning that aims to solve several NLP tasks without the need of fine tuning the model on a task specific dataset. Transformer language models have been shown to increase performance based on the amount of parameters used during training [20], leading to GPT 3 being trained with 175 billion parameters, a substantial amount compared to BERT's 110 million.

GPT 3 is a decoder based language model that works by queries (also called prompts) given by a user. The model processes the user's query and generates an adequate response, but the response varies based on the structure of the query. GPT 3 was trained using three types of queries: Zero-shot, where there's no example of the intended use of the query; One-shot, where the query shows just one example of the task; and Few-shot, where the query shows multiple examples of the task in hand.

For our task we focus on testing GPT 3.5's Zero-shot behavior in the context of Question Answering, this means that user prompts given to the model will be solely the questions with which we aim to evaluate. The only adjustment made will be a system prompt that modifies GPT 3.5 behavior for all iterations.

As we can see, the architecture, dataset, and evaluations are very similar in the models for answer generation. The difference lies in their size and whether they are open access or paid. Additionally, in the discussion section, we will mention the differences in the models' responses. It's worth noting that for GPT, in it's 3.5 version, the related article does not provide sufficient information about the architecture or the dataset.

4. Corpus and Questions

The Bible stands as the most widely read text in Western history. Additionally, its narrative ranks among the most renowned and globally significant with a huge number of translations. Consequently, many datasets include this collection of texts in their repository; this means most LLMs should have knowledge of biblical writings. The Bible is also employed as a tool for the creation of parallel corpora [21] for automatic translation. It is also frequently used as a reference corpus for paraphrase detection [22] since some languages have several versions and translations of the text. With this in mind, we consider The Bible a good option for evaluating the LLMs mentioned above, taking into account the accessibility of the text.

The digital edition we selected was *La Biblia - Latinoamérica* [23], taken from the corpus created by Sierra *et al.* [24]. This version has labels for each verse, which facilitates the organization of the corpus since the information is clearly separated. Also, each verse is manually annotated and revised to avoid any misleading information. This structure is computationally efficient, particularly for NLP techniques, since no additional segmentation, has to be carried out for most tasks.

To evaluate the performance and behaviour of the LLMs, we wrote a series of question about The Bible, some of the questions can be answered based on a specific passage of The Bible and some can not. In total, 36 questions were selected for the final version of the experiment. The majority of questions are about specific events, names, moments, and places. In general, the information necessary to ask the question is available entirely from one chapter to facilitate the retrieval of information. These questions were created in this manner in order to limit each LLM's answer to factual information only. The focus of the experiment is to test an LLM's factual information reasoning, not religious interpretation of this text. These questions require the LLMs to consider The Bible as a source of information, not as religious guidelines or as information that contradicts the model's pretrained knowledge.

The questions formulated for the examination were categorized into two distinct groups:

1. Those that could be addressed using solely the information furnished by The Bible (Table 1).
2. Those that necessitated additional information beyond that provided by the scriptural text for resolution (Table 2).

 doi:10.20944/preprints202411.0502.v1

Table 1. Questions that can be answered with information contained in The Bible.

#	Question	Expected answer
1	¿Qué mar fue abierto por Moisés y para qué? *Which sea was parted by Moses, and for what purpose?*	El mar Rojo, to escape from the Egyptians. *The Red Sea, .*
2	¿Qué ídolo erróneamente veneran los israelitas? *What idol did the Israelites mistakenly worship?*	Un becerro de oro. *A golden calf.*
3	¿Cómo mató David a Goliat? *How did David kill Goliath?*	Con una piedra de su honda. *With a stone from his sling.*
4	¿Quién tuvo el sueño de las vacas gordas y las vacas flacas? *Who had the dream of fat cows and lean cows?*	El faraón de Egipto. *The Pharaoh of Egypt.*
5	¿A quién le dijo Rut las palabras: "donde tú vayas, iré yo; y donde tú vivas, viviré yo; tu pueblo será mi pueblo y tu Dios será mi Dios"? *To whom did Ruth say the words: "Where you go, I will go; and where you stay, I will stay. Your people will be my people and your God my God"?*	A su suegra Noemí. *To her mother-in-law Naomi.*
6	¿Cómo se llamaba el jefe del ejército a quien derrotaron los israelitas bajo el mando de la jueza Débora? *What was the name of the army chief defeated by the Israelites under the command of the judge Deborah?*	Sísera. *Sisera.*
7	¿Quién mató a Holofernes? ¿Cómo? *Who killed Holofernes and how?*	Judith. Lo decapitó. *Judith. She decapitated him.*
8	Who stripped Samson of his hair and why? *Who stripped Samson of his hair and why?*	Su esposa, Dalila, para que perdiera su fuerza. *His wife, Delilah, so that he would lose his strength*
9	¿Qué comían los israelitas en el desierto? *What did the Israelites eat in the desert?*	Maná. *Manna.*
10	¿Quién era Nabucodonosor? *Who was Nebuchadnezzar?*	El rey de Babilonia o Asiria. *The king of Babylon.*
11	¿Qué oficio tenía Melquisedec? *What profession did Melchizedec have?*	Sacerdote. *Priest.*
12	¿Qué construyó Noé? ¿De qué escapaba? *What did Noah build? What was he escaping from?*	Un arca. De un diluvio. *An ark.From a flood.*
13	¿De dónde era Ciro? *Where was Cyrus from?*	Ciro era de Persia. *Cyrus was from Persia.*
14	¿Quién tentó a Jesús en el desierto? *Who tempted Jesus in the desert?*	Satanás. *Satan.*
15	¿En qué monte fue crucificado Jesús? *On what mount was Jesus crucified?*	Gólgota. También llamado Calvario. *Golgotha. Also called Calvary.*
16	¿Qué pareja acompañó a Pablo en algunos de sus viajes? *Which couple accompanied Paul on some of his travels?*	Aquila y Priscila. *Aquila and Priscilla.*
17	¿Quién se quedó sin oreja la noche que murió el maestro? *Who lost his ear the night that Jesus died?*	Malco. *Malchus.*
18	¿Quién estaba siendo juzgado junto con Jesús por los romanos? *Who was being tried alongside Jesus by the Romans?*	Barrabás. *Barabbas.*
19	¿Qué le hizo Juan Bautista a Jesús? *What did John the Baptist do to Jesus?*	Lo bautizó. *He baptized him.*
20	¿Por cuántas monedas Judas traiciona a Jesús? *How many coins did Judas betray Jesus for?*	Treinta piezas de monedas de plata. *Thirty pieces of silver coins.*
21	¿Quién hizo que decapitaran a Juan Bautista? *Who caused John the Baptist to be beheaded?*	La hija de Herodías.. *The daughter of Herodias.*
22	¿Qué amigo le escribe dos cartas a Timoteo? *Which friend wrote two letters to Timoteo?*	Pablo de Tarso. *Pablo de Tarso.*
23	¿Quién niega a Jesús? ¿Cuántas veces? *Who denies Jesus? How many times?*	Pedro. Tres. *Peter. Three times*
24	¿A qué hora murió Jesús? *At what time did Jesus die?*	A las tres de la tarde. *At three in the afternoon.*

 doi:10.20944/preprints202411.0502.v1

Table 1. *Cont.*

25	¿En qué ciudad hizo Pablo de Tarso su discurso "Al Dios desconocido"?	Atenas.
	In what city did Paul of Tarso give his speech "To the Hidden God"?	*Athens.*
26	¿En qué fueron grabados los diez mandamientos y cuáles son esos?	Fueron dados en dos tablas de piedra y son: 1. Amarás a Dios sobre todas las cosas. Sólo existe un Dios, creador y todopoderoso, al que adorar. 2. No tomarás el nombre de Dios en vano. 3. Santificarás las fiestas. 4. Honrarás a tu padre y a tu madre. 5. No matarás. 6. No cometerás actos impuros. 7. No robarás. 8. No darás falso testimonio ni mentirás. 9. No consentirás pensamientos ni deseos impuros. 10. No codiciarás los bienes ajenos.
	On what were the Ten Commandments engraved, and what are they?	They are: 1. You shall love God above all things. There is only one God, creator and almighty, to whom worship. 2. You shall not take the name of God in vain. 3. You shall keep holy the feast days. 4. You shall honor your father and your mother. 5. You shall not kill. 6. You shall not commit impure acts. 7. You shall not steal. 8. You shall not bear false witness nor lie. 9. You shall not consent impure thoughts or desires. 10. You shall not covet your neighbor's goods.
27	¿Cuántos hijos tuvo Jacob? ¿Cómo se llamaban?	12.Rubén, Simeón, Leví, Judá, Dan, Neftalí, Gad, Aser, Isacar, Zabulón, José, Benjamín.
	How many sons did Jacob have? What were their names?	12. Reuben, Simeon, Levi, Judah, Dan, Naphtali, Gad, Asher, Issachar, Zebulun, Joseph, Benjamin.
28	¿Cuál era el oficio de Mateo antes de unirse a los seguidores de Jesús? ¿Y de Pedro?	Recaudador de impuestos (publicano), pescador.
	What was Matthew's occupation before joining Jesus' followers? And Peter's?	*Tax collector (publican). Fisherman.*
29	¿Cuántos candeleros hay en Apocalipsis y a qué se refiere?	Siete. A las 7 iglesias.
	How many lampstands are there in Revelation, and what do they refer to?	*Seven. To the seven churches..*
30	¿A quién se tragó el pez grande?	Jonás.
	Who was swallowed by big fish?	*Jonah.*
31	¿A quién le fue revelado el libro del Apocalipsis?	A Juan.
	To whom was the book of the Apocalypses revealed?	To John.

The initial category (question group 1) of questions typically proves simpler to answer as the required responses are directly available within a single, or multiple chapters of The Bible.

Conversely, to answer the second type of questions (question group 2), more information than that provided by the context is required. In response to these questions, it is expected that LLMs may not be able to provide answers, since the necessary information is not explicitly presented in the books of The Bible. This information usually consists of interpretations and opinions formulated by scholars or ecclesiastical authorities.

As can be seen in the questionnaire, questions focus on particular individuals with particular occupations, locations, and actions. These details are expressed purely with lexical elements (the names of the individuals, actions, etc). Since dense embeddings work with (varying) context windows, we run the risk of mixing and losing certain named entities given that they can all be clustered in a

 doi:10.20944/preprints202411.0502.v1

biblical category. Being able to use lexical embeddings, or a weighted sum that takes this lexical factor into consideration, might prove to be a better retrieval method for very precise queries.

Table 2. Questions that need more information than the one provided by the text to be resolved.

#	Question	Expected answer
32	¿Qué libro de la Biblia narra el amor de los esposos? *Which book of The Bible tells the love of spouses?*	El Cantar de los Cantares.. *The Song of Songs.*
33	¿Quién es considerado el autor de los salmos? *Who is considered the author of the Psalms?*	El rey David. *King David.*
34	¿Qué son Isaías, Jeremías, Ezequiel y Daniel? *What are Isaiah, Jeremiah, Ezekiel, and Daniel?*	Profetas. Los profetas mayores. *Prophets. The major prophets.*
35	¿Qué profeta escribió el libro de las Lamentaciones? *Which prophet wrote the book of Lamentations?*	Jeremías. *Jeremiah.*
36	¿Cuál era el más escéptico de los discípulos de Jesús? *Who was the most skeptical of Jesus' disciples?*	Tomás. *Thomas.*

5. Methodology/RAG

In this section the precise workflow of the system is described. In order to correctly implement the aforementioned LLMs, corpus, and questionnaires, a system capable of integrating these components simultaneously, and in a standardized and comparable manner is necessary. As one might guess a Retrieval Augmented Generation (RAG) [8] approach was elected, opting to explore two of the original RAG papers' question answering tasks: Open-domain Question Answering and Abstractive Question Answering.

These experiments evaluate the LLMs' question answering behavior and the question answering itself. What we mean by behavior is whether the model sticks only to the extracted information and reasons based solely on that information. Whereas question answering corresponds to the actual answer given by the model. It's important to observe that a correct answer should only be obtained from a correct information extraction. If the model bypasses the extracted information and uses prior knowledge, then the behavior is incorrect.

In accordance with this, certain questionnaire questions are designed to require a deeper degree of understanding and rationalization rather than just extracting the correct answer from the context. For some questions the presented context will never generate a correct answer, nonetheless a context has been extracted and used in order to evaluate behavior in this scenario. In the following section evaluation will be described in detail.

In order to generate a consistent ground for comparison and removing noise generated from each model's retrieval process, the retrieval aspect of the model was done manually with standardized embeddings. Questions and chapters were embedded using the same embedding model after which a similarity retrieval was carried out.Using the obtained texts, each LLM was asked to use the retrieved context to answer the corresponding question. Figure 1 shows the workflow of the RAG system.

The rest of this section will describe the previous paragraph and diagram in detail, starting with an analysis of the dataset, followed by the text preprocessing, the embedding creation, the embedding evaluation, the answer generation and the answer evaluation.

5.1. Analysis of the Dataset

We have chosen for our experiments the texts from *La Biblia Latinoamericana*, this translation of The Bible includes 73 books, with a total of 1,326 chapters, 35,245 verses, and 768,282 words. The length of each chapter ranges anywhere from 13 to 2089 words, averaging 579 words per chapter, with a standard deviation σ(verse)=295. This presents problems since it indicates that the chapters vary heavily in length meaning that any embedding generation has to be able to work around this. In contrast, the questionnaire has 36 questions whose length ranges from 3 to 29 words, and with σ(questions) = 9, making them more suitable to almost any embedding model.

 doi:10.20944/preprints202411.0502.v1

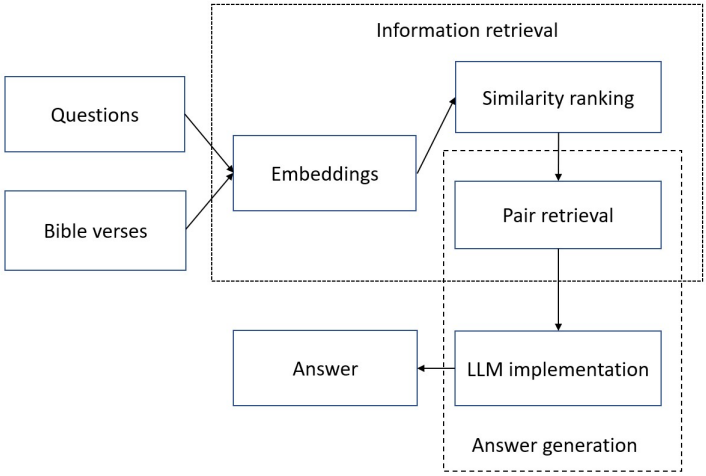

Figure 1. The RAG system flow of information. The questions and Bible chapters correspond to the input of the system, whereas the LLM's answer corresponds to the output.

These statistical measures regarding the length of texts are highly relevant when working with LLMs. Token limits for language and embedding models are an ever-present problem which has to be taken into consideration. Understanding these values allows an adequate segmentation of texts and correct LLM usage during the experimental setup. Afterwards, it helps the evaluation of the system itself since details like tokens and segmentation lengths can be thought of as modifiable hyper-parameters.

5.2. Text Pre-Processing

The objective of the experiments is to test LLMs' capabilities of reasoning over texts of varying length based on questions of varying complexity, analyzing both the computational process and the answers themselves. Arbitrarily truncating biblical chapters from which the answer of a question has to be *reasoned* is a counterproductive measure, resulting in it not being carried out. In a similar manner, since LLMs are capable of working with natural language text in a crude form, there's no need to tokenize, lemmatize, or modify the text or the questions, resulting in completely unmodified texts from the dataset to the model.

5.3. Embeddings and Retrieval

Given the high variance of chapter length, as well as the fact that they're extracted from a Spanish translation of The Bible, the embedding model needed to generate adequate vectors has to be available to work with various elements, mainly: high token count and multilinguality. The BGE-M3 Embedding project suits these necessities in an optimal manner. The following paragraphs dive deeper into BGE-M3's functionality.

As mentioned before, these embeddings work considering a three-fold granularity meaning they maintain performance with sentence, passage, and document level embeddings. In addition, BGE-M3 Embeddings are capable of carrying out a similarity search over their three type of embeddings simultaneously, meaning that lexical, dense, and multi-vector retrieval are combined into one hybrid step, maintaining various key information retrieval aspects present during the extraction.

The chapters and the questions were embedded using a 8192 token maximum length, and with a batch size of 12 and the three different types of embeddings, dense, lexical and multi-vector, were computed. Once the embeddings were ready, the vector comparison and subsequent information retrieval section was carried out using BGE-M3's similarity metrics for each type of embedding. For each element of the questionnaire the most similar chapter was selected based on these metrics for the three different types of embeddings and for its combinations.

 doi:10.20944/preprints202411.0502.v1

The result are 6 different sets of 36 ordered pairs of question and chapters each one of them corresponding to the comparison of each type of embedding, the combinations and the weighted combination. For the weighted combination, we use the following weights, $[0.4, 0.2, 0.4]$, for dense, lexical and multi-vec scores, respectively. After an evaluation, we selected one set to be used during the answer generation part of the system.

5.4. Generation

After the information retrieval portion of the system, we're left with the evaluation of the questionnaire based on the retrieved context. The aim of these experiments is to a) analyze if LLMs are able to restrict themselves to the presented information and prompts, b) if they disregard the command and the context given in order to correctly answer a question using their prior knowledge, and c) if the given information modifies a potentially correct answer or behavior. In order to fully measure these objectives two evaluations were carried out.

The first evaluation consists of simply asking the LLM each of the questions in the questionnaire without providing any further instructions, it was only to respond that it does not have the information unless it is certain of the answer. Each answer will correspond to a neutral behavior and will originate solely from each LLM's prior knowledge. This means that this first evaluation does not use the retrieval system presented in Section 5.3, and will serve as a standard to compare the complete RAG system. Systems' prompt are depicted in Table 3.

Table 3. Prompt used as LLMs' neutral behavior.

Prompt
Responde la pregunta siguiente. Si no estás seguro de la respuesta, por favor responde "No tengo dicha información". Pregunta: {query} Respuesta:
Answer the following question. *If you are not sure of the answer, please respond with 'I do not have that information.'.* Question: {query} Answer:

The second evaluation however does, in fact, use the retrieved text and question pairs obtained in Section 5.3. Various prompts were tested for each model, and as much as the precise structure varied, the underlying request remained the same. This structure asks the model to use the chapter given as context and then asks the model to answer the question solely using the context given. If the context proves to be insufficient when it comes to answering the question, then the model should answer with a comment that indicates this fact. Table 4 indicates the used prompt for each model. An approximate translation of the general idea of the prompt is: Answer the question given the following context. If the answer is not within the context, and if you're unsure of it answer with "Information not available".

Each question and context was passed to the model in an iterative manner, returning a single answer per iteration. Each answer was collected and stored for evaluation alongside the corresponding question and context, this due to the fact that each LLM's behavior will be measured based on these 3 elements. The result of this whole process yields a 3-tuple: $(Question, Context, Answer)$. Once the 36 3-tuples are obtained, the question by question analysis is carried out. Results are presented in the following section.

 doi:10.20944/preprints202411.0502.v1

Table 4. Prompt with RAG implementation.

Prompt
Responde la pregunta dado el siguiente contexto: {{context:}} Si la respuesta no está dentro del {{context:}} y no estás seguro de la respuesta, por favor responde "Información no disponible en el contexto". Contexto: {retrieved_info} Pregunta: {query} Respuesta:
Answer the question given the following context: {{context:}} *If the answer is not within the {{context:}} and you are not sure of the answer, please respond with 'Information not available in the context'.* *Context*: {retrieved_info} *Question*: {query} *Answer*:

6. Results

The following section presents the results of the RAG system in a traditional question answering manner, results are divided in the retrieval and generation section. The former will briefly analyze the results of BGE-M3's hybrid search similarity, regarding the chapter search part of the retrieval. The latter will compare each generated answer for Llama 2 Chat, GPT 3.5 and PaLM models, with the gold standard provided in the dataset.

6.1. Retrieval Results

As explained in the methodology, we obtained 6 sets using the three different types of embeddings computed by BGE-M3 model. The results were grouped by type of question and embedding and are showed in Table 5.

Table 5. IR Embeddings Evaluation

Question Group	Correct	Incorrect
1	19	12
2	0	5

(a) Dense

Question Group	Correct	Incorrect
1	16	15
2	0	5

(b) Lexical

Question Group	Correct	Incorrect
1	18	13
2	1	4

(c) Multi-Vec

Question Group	Correct	Incorrect
1	20	11
2	0	5

(d) Dense+Lexical

Question Group	Correct	Incorrect
1	20	11
2	1	4

(e) Dense+Lexical+Multi-Vec

Question Group	Correct	Incorrect
1	22	9
2	1	4

(f) Dense+Lexical+Multi-Vec Weighted

The weighted combination of the three distinct similarity scores computed by BGE-M3 yielded the highest performance in the system's retrieval component, identifying 23 correct passages. Notably, one of these passages corresponds to the second group of questions, which, in theory, cannot be answered solely using the content within The Bible. In Table 6, the column "Chapter" shows G (Good) if the passage retrieved is correct for the given question and B (Bad) if it is incorrect.

6.2. Generation Results

In the case of LLMs, we evaluate first the RAG system, where we provide the passage retrieved, the question alongside the prompt depicted in Table 4 to the LLM in order to restrict the answer

 doi:10.20944/preprints202411.0502.v1

generation over the given passage. In this case we evaluate as Good (G), if the behaviour of the LLMs is what we expected to be, this means that the LLM does what it was told to do in the prompt. In addition, we marked those questions where the models indicated that there was no information in the given context as "G/Expected".

In Table 6 we present the evaluation for the RAG systems.
For the answer evaluation using solely the LLMs' prior knowledge the results are in the Table 7, question where the model answer "I do not have that information" are marked as "G/Expected".

Table 6. Answer Evaluation using RAG.

Question	Chapter	Llama 2 Chat	GPT 3.5	PaLM
1	G	G	G	G
2	G	G	G/Expected	G
3	G	G	G	G
4	G	G	G	G
5	B	G/Expected	G/Expected	G/Expected
6	G	G	G	G
7	G	G	G	G
8	G	G	G	G
9	G	G	G	G
10	G	G	G	G
11	G	G	G	G
12	G	G	G	G
13	G	G	G	G
14	G	G	G	G
15	B	G/Expected	G/Expected	G/Expected
16	B	B	B	B
17	B	G	G/Expected	G
18	G	B	B	G
19	B	G/Expected	G/Expected	G/Expected
20	G	B	G	G
21	G	B	G	G
22	B	G/Cheat	G/Expected	G/Cheat
23	G	G	G	G
24	B	G/Expected	G/Expected	G/Expected
25	G	G	G	G
26	B	G/Cheat	G/Expected	G/Expected
27	B	B/Incomplete	G/Cheat	G/Cheat
28	G	G	B	B
29	G	G	G	G/Incomplete
30	G	G	G	G
31	G	G	G	G
32	B	G/Expected	B	B
33	B	B	G/Expected	G/Cheat
34	B	G/Cheat	G/Expected	G/Expected
35	B	G	G/Expected	G
36	G	G	B	G

Regarding the evaluation of the LLMs, it can be seen that if the correct chapter was retrieved and presented, then most of the answers are correct. Similarly, there is a pattern between an incorrect chapter and incorrect answers, meaning that a lack of information leads the model to answering poorly.

When a model disregards the chapter provided as context and relies on its pretrained data to accurately answer a question, it is deemed to be exhibiting a form of cheating. LLaMA 2 chat exhibits this behavior in three instances, GPT-3.5 in one instance, and PaLM in three instances. It is important to note that the questions in which the models engaged in this deceptive practice involved incorrect retrievals; however, the anticipated behavior would have been for the models to respond with "Information not available in the context.

 doi:10.20944/preprints202411.0502.v1

Table 8 shows a comparison in numbers between using and not using the RAG system by model. Analysis of all of these results is shown in the next section.

Table 7. Answer Evaluation without RAG.

Question	Llama 2 Chat	GPT 3.5	PaLM
1	G	G/Expected	G/Incomplete
2	G/Expected	G/Expected	G/Incomplete
3	G	G	G/Incomplete
4	G/Expected	G/Expected	G/Expected
5	G	G	G/Expected
6	G	G	G
7	G	G	G/Incomplete
8	G	G	G/Incomplete
9	B	G	G
10	G	G	G
11	G	G/Expected	G
12	G	G/Expected	G/Expected
13	B	G/Expected	G
14	B	G	G
15	G/Expected	G	G
16	G/Expected	B	G
17	G/Expected	G/Expected	B
18	G/Expected	G	G
19	G	G/Expected	G
20	G	G	G
21	B	G	G
22	G	G	G
23	G/Expected	G/Expected	G
24	G/Expected	G/Expected	G
25	G/Expected	G	G
26	G	G	G/Expected
27	G	G	G/Incomplete
28	G	G	G/Expected
29	G/Expected	G/Expected	G/Incomplete
30	G/Expected	G/Expected	G
31	G	G	G
32	G	G	G
33	G	G	G
34	G	G	G
35	G/Expected	B	G
36	G/Expected	G/Expected	G

 doi:10.20944/preprints202411.0502.v1

Table 8. Evaluation RAG vs No RAG.

	RAG	No RAG
Correct	23	19
Incorrect	5	4
No data	5	13
Cheat	3	N/A

(a) Llama 2 Chat

	RAG	No RAG
Correct	19	21
Incorrect	5	2
No data	11	13
Cheat	1	N/A

(b) GPT 3.5

	RAG	No RAG
Correct	24	30
Incorrect	3	1
No data	6	5
Cheat	3	N/A

(c) PaLM

7. Discussion

The result discussion considers two different aspects: retrieval and answer generation. Afterward, an analysis of solutions provided by the system will be carried out. In this study, we do not consider F1 or Accuracy measures for quantitative evaluation. Furthermore, the aspects we aimed to assess, the performance and behavior of the models, cannot be evaluated using these metrics.

In the retrieval part, the weighted combination of the three different embedding vectors that BGE-M3 creates is able to retrieve 23 passages containing chapters with relevant information, out of 36 possible.

It's worth noting that this system is capable of retrieving a relevant chapter for question 36. This question is one of the five questions marked as unanswerable (question group 2) since the answer requires information that goes beyond the scope of the text. This suggests that BGE-M3's model is capable of recognizing relevant information even in truncated scenarios.

Although the chapters themselves doesn't contain the exact answer, they contain something relevant to the question. The similarities might be lexical (lexical overlap in Bible chapters isn't rare), or in the mathematical representation of the embedding itself, regardless this embedding model is capable of identifying them and pairing them, most of the time accordingly.

Nonetheless this can be misleading, as shown in question number 2: "¿Qué ídolo erróneamente veneran los israelitas?" (*What idol did the Israelites mistakenly worship?*) The answer is a golden calf. The retrieved chapter, Isaiah 44, mentions something about false gods made out of wood, as well as the Israelis, but it doesn't contain anything about a golden calf. This embedding model is susceptible of interpreting wood as the key element of the chapter, and prioritizes an incorrect chapter in this context. We marked the answer for this question as Good for Llama and PaLM because the models' answer was based on the retrieved context.

This also occurs in question number 17: "¿Quién se quedó sin oreja la noche que murió el maestro?" (*Who lost his ear the night that Jesus died?*). The answer is "Malchus". Our system retrieved, Marcus 14, which describes the night on which Jesus died and the incident in which "the humble servant of the highest priest loses an ear"; however, it does not mention his name, whereas in John 18 besides the previous description the name Malco is mentioned. Information discrepancies like this affect the LLM's answer in an unwanted way. We marked the answer for this question as Good for Llama and PaLM also because the models' answer was based on the retrieved context.

 doi:10.20944/preprints202411.0502.v1

These questions are prone to being answered poorly by the model since an adequate retrieval of a chapter still leaves out relevant information. As an example we can see question number 27: "¿Cuántos hijos tuvo Jacob? ¿Cómo se llamaban?" *How many sons did Jacob have? What were their names?*. The chapter that the system retrieved only mentions the first four of Jacob's children, in this case Llama 2 Chat answers only the four mentioned childrens given in the context, meanwhile GPT 3.5 an PaLM resort to their prior knowledge to answer correctly, that is why we marked as cheat.

As a final example regarding retrieval, in question 36, "¿Cuál era el más escéptico de los discípulos de Jesús¿'(*Who was the most skeptical of Jesus' disciples?*), which belonged to the second group, the system retrieve John 20 that states the skepticism of Thomas, despite the existence of other chapters in The Bible that depict followers of Jesus expressing skepticism. For instance, this includes Peter when he denies Jesus three times, or when he hesitates to walk on water.

Furthermore, there were questions where the retrieved chapter was correct; nonetheless, the models were unable to infer the response based on it. Llama 2 Chat performed this in questions 18, 20, and 21, GPT 3.5 in questions 18, 28, and 36, and PaLM only in question 28.

Regarding answer generation we can observe that there is a significant difference in the size and amount of vocabulary used by the different models. Naturally this can be explained by the nature of the models used. We use Llama 2 Chat, which is optimized for dialogue use cases, unlike PaLM, whose base model is just for text generation.

We also observed that some of the responses provided by LLama 2 Chat, characterized by its neutral behavior, were contradictory. Specifically, it initially provided the correct answer but subsequently stated that it did not possess the information.

From Table 6 we can observe that providing the model with a relevant chapter to answer the question doesn't necessarily improve the total amount of correct answers. Llama improved from 19 to 23 correct answers. Surprisingly GPT 3.5 and PaLM, provided with the correct chapter, do not improve their correct answers. GPT 3.5 without chapter information has 21 correct answers whereas with chapter information the correct answers go down to 19. In the other hand, PaLM without chapter information has 30 correct answers, however, 7 of those responses were incomplete, and 24 with chapter information.

8. Conclusions

In this work, we employ the RAG methodology to leverage the prior knowledge and capabilities of large language models, specifically Llama 2 Chat, GPT 3.5, and PaLM.

One key insight from this study is the demonstration of these models' abilities to respond to questions based on a given context. The variance in correct answers between models with and without context relies on factors such as information retrieval, model size, and data availability. This inconsistency is one of the reasons why precision and accuracy measures are not particularly helpful in this type of task and it is recommended to conduct a different evaluation.

It should be noted that we cannot guarantee the absence of The Bible in the pretraining data of any of the three models and it is likely that this explains why the models tend to rely on their prior knowledge instead of following the given instructions in some of the questions.

This methodology facilitates easy domain adaptation, and with constant model upgrades, it can be utilized without the need for fine-tuning. As a future work

As future work, we aim to test newer models, as well as implement the use of multiple passages within the models, since we observed that some questions require more than one passage or necessitate inferences from multiple passages. Additionally, we plan to utilize a corpus that we can confirm does not belong to the model's prior knowledge to determine if this approach allows the model to behave according to the provided instructions.

It's important to mention the significance of having open-source models since the computational and economic cost of generating such models is only accessible to a few of the largest companies. With the RAG methodology, we can leverage these systems to avoid training a model of this size.

 doi:10.20944/preprints202411.0502.v1

Author Contributions: Conceptualization, B-E.G. and S.G.; methodology, CB.A. and LP.F.; experiments CB.A. and LP.F.; validation, B-E.G. and S.G and O-T.S.; formal analysis, CB.A. and LP.F.; data curation, O-T.S.; writing—original draft preparation, CB.A. and LP.F.; writing—review and editing, B-E.G. and S.G.; supervision, B-E.G. and S.G.;funding acquisition, B-E.G. and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CONAHCYT (CF-2023-G-64) and PAPIIT (IT100822).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Louis, A.; van Dijck, G.; Spanakis, G. Interpretable long-form legal question answering with retrieval-augmented large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 22266–22275.
2. Tan, Y.; Zhang, Z.; Li, M.; Pan, F.; Duan, H.; Huang, Z.; Deng, H.; Yu, Z.; Yang, C.; Shen, G.; others. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Computers in Biology and Medicine* **2024**, *172*, 108290.
3. Lakkaraju, K.; Jones, S.E.; Vuruma, S.K.R.; Pallagani, V.; Muppasani, B.C.; Srivastava, B. LLMs for Financial Advisement: A Fairness and Efficacy Study in Personal Decision Making. Proceedings of the Fourth ACM International Conference on AI in Finance, 2023, pp. 100–107.
4. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* **2024**, *55*, 1–38. doi:10.1145/3571730.
5. Ángel Cadena.; López-Ponce, F.; Sierra, G.; Lázaro, J.; Ojeda-Trueba, S.L. Information Retrieval Techniques for Question Answering based on Pre-Trained Language Models. *Research in Computing Science* **2023**, *152*.
6. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. REALM: Retrieval-Augmented Language Model Pre-Training, 2020, [arXiv:cs.CL/2002.08909].
7. Lee, K.; Chang, M.W.; Toutanova, K. Latent Retrieval for Weakly Supervised Open Domain Question Answering. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Korhonen, A.; Traum, D.; Màrquez, L., Eds.; Association for Computational Linguistics: Florence, Italy, 2019; pp. 6086–6096. doi:10.18653/v1/P19-1612.
8. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2021, [arXiv:cs.CL/2005.11401].
9. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; Merlo, P.; Tiedemann, J.; Tsarfaty, R., Eds.; Association for Computational Linguistics: Online, 2021; pp. 874–880. doi:10.18653/v1/2021.eacl-main.74.
10. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024, [arXiv:cs.CL/2312.10997].
11. Li, X.; Liu, M.; Gao, S. GRAMMAR: Grounded and Modular Methodology for Assessment of Domain-Specific Retrieval-Augmented Language Model, 2024, [arXiv:cs.CL/2404.19232].
12. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine, 2024, [arXiv:cs.CL/2402.13178].
13. Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; Fleisch, B. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering, 2024, [arXiv:cs.CL/2404.04302].
14. Zhao, H.J.; Liu, J. Finding Answers from the Word of God: Domain Adaptation for Neural Networks in Biblical Question Answering. 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8. doi:10.1109/IJCNN.2018.8489756.
15. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024, [arXiv:cs.CL/2402.03216].
16. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C.C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P.S.; Lachaux, M.A.; Lavril, T.; Lee, J.; Liskovich,

 doi:10.20944/preprints202411.0502.v1

D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E.M.; Subramanian, R.; Tan, X.E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J.X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023, [arXiv:cs.CL/2307.09288].

17. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models are Few-Shot Learners, 2020, [arXiv:cs.CL/2005.14165].
18. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A.M.; Pillai, T.S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; Fiedel, N. PaLM: Scaling Language Modeling with Pathways, 2022, [arXiv:cs.CL/2204.02311].
19. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models, 2023, [arXiv:cs.CL/2302.13971].
20. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *CoRR* **2020**, *abs/2001.08361*, [2001.08361].
21. Sierra, G.; Montaño, C.; Bel-Enguix, G.; Córdova, D.; Mota Montoya, M. CPLM, a Parallel Corpus for Mexican Languages: Development and Interface. Proceedings of the Twelfth Language Resources and Evaluation Conference; Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S., Eds.; European Language Resources Association: Marseille, France, 2020; pp. 2947–2952.
22. Ponce, F.F.L.; Sierra, G.; Enguix, G.B. Sistemas de clasificación aplicados a la detección de paráfrasis.
23. Pablo, E.S., Ed. *La Biblia, Latinoamérica*; 1989; p. 1728. Estimated reading time: 41h 38m.
24. Sierra, G.; Bel-Enguix, G.; Díaz-Velasco, A.; Guerrero-Cerón, N.; Bel, N. An aligned corpus of Spanish bibles. *Language Resources and Evaluation* **2024**, pp. 1–31.