Document-based question answering

Hachem Betrouni[1,2]

[1]National Polytechnic school of Algiers, Industrial Engineering Department, Data Science and AI, Algiers, hachem.betrouni@g.enp.edu.dz
[2]BIGmama technology, France, hb@big-mama.io

June 19, 2023

Abstract

In this thesis, we present a comprehensive analysis of various open-source embedding models and large language models (LLMs) for document-based question answering (DBQA). The goal of DBQA is to extract relevant information from a given document and answer user queries in natural language. We evaluate the performance of different embedding models, including BERT, E5, and MiniLM, in combination with open source LLMs such as Vicuna, Falcon, and OpenLlama. Our experimental setup involves a retriever-generator framework, where a retrieval system retrieves the most relevant contexts from a document using embeddings of document and query. Then we condition the generative system (LLM) with the most promising context to generate a response to the user. We benchmark the performance of these systems using a sample from the SQuAD dataset and GPT-3 as a judge. We find that the choice of embedding models significantly impacts the retrieval performance. Moreover, we observe that the choice of LLM also plays a crucial role in generating accurate and informative answers. This comparative analysis provides insights into the strengths and weaknesses of various open-source embedding models and LLMs for DBQA tasks. The findings of this study can guide researchers and practitioners in selecting suitable models for their specific DBQA applications and contribute to the advancement of question-answering systems.

To my parents,
to my family, mentors, friends, enemies, and every person who shaped who I am,
to every rock that broke my bones, to every sea waving on the shores,
to Algiers's seagulls waking me up morning, reminded of Allah's boons,
to Collo's men who rescued me,
to you all I dedicate not this thesis but the years that led to its maketh.

Contents

List of Figures

List of Tables

1 General introduction

For this BIGmama confined the development of a document-based question answering model to us.

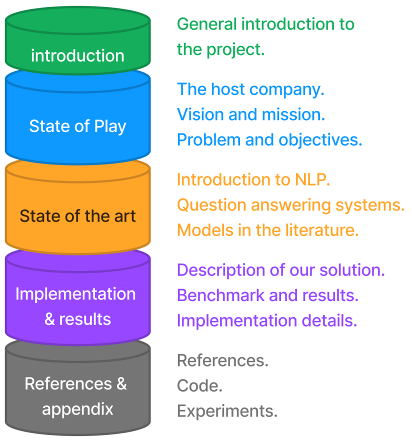

Figure 1: General structure of this document.

2 State of play

2.1 Introduction

Creating value in the market is usually about solving a complex problem or answering a question. When it comes to the Data science and AI sector this used to be tackled with off-the-shelf solutions/applications/answers. However the arrival of LLMs and other fundamental models upset the level of requirements. Creating value now requires expensive, bespoke AI solutions. The requirements are no longer to predict a number, but to explain the prediction, it's no longer to recommend a blog title but to generate its whole content.

2.2 Asset-based consulting

This evolvement of the business landscape, driven by rapid technological advancements, changing market dynamics, and increasing competition created a fast-paced environment, where organizations seek innovative ways to optimize their operations, unlock hidden potentials, and achieve sustainable growth.

Today we are well aware that answering these organizations needs for innovation has shifted from writing consulting reports (or any of the traditional approaches) to building tangible assets that concretize the added value. Today, clients are expecting specialised solutions to their problems, rather than the boilerplate solutions consultants traditionally provided.

This has given rise to a thriving industry known as asset-based consulting (ABC).

Asset-based consulting is a specialized field within consulting that focuses on creating, leveraging, and maximizing the value of a company's assets, AI can be incorporated in the schema of consulting by leveraging the data/knowledge/expertise of the company, incorporating it with intelligent systems, automatize tasks and improve the productivity of its users in general.

In the ABC context, assets can include tangible resources such as infrastructure, technology, and inventory, as well as intangible assets like intellectual property, brand equity, and human capital. By harnessing the full potential of these assets, asset-based consultants help organizations drive operational efficiency, enhance performance, and create a competitive advantage.

Unlike traditional consulting approaches that often emphasize external factors and strategies, asset-based consulting takes a more holistic and internal perspective. It recognizes that companies possess unique assets and capabilities that, when strategically managed, or in our case, combined with AI, can serve as the foundation for sustainable competitive advantage and resilience. Here consultants work closely with clients to understand their expertise, identify untapped questions and needs, and design tailored AI solutions.

Asset-based consulting is known for its multidisciplinary nature, drawing expertise from various fields including finance, operations, technology, marketing, and human resources.

Consultants blend their industry knowledge, analytical skills, and business acumen to assess, diagnose, and unlock the hidden value with innovative assets. They often conduct thorough assessments, employ data-driven methodologies, and collaborate closely with expert teams to align the asset usefulness with the overall business objectives.

As industries become increasingly complex and competitive, the demand for asset-based consulting continues to rise. Organizations of all sizes and sectors recognize the need to unlock the value with AI assets to stay ahead of the curve. From manufacturing firms seeking to optimize their supply chains to technology companies aiming to capitalize on their intellectual property, asset-based consulting provides a strategic framework and expertise to achieve these objectives.

2.3 BIGmama technology

BIGmama is a French registered startup with an Algerian extension, specialized in data science and AI. They have been developing bespoke predictive applications for more than 8 years (as of the year 2023).

The board of directors counts former CEOs of global groups (Danone, Safran) and the team is assembled from a dozen of high-level data scientists and software engineers.

2.4 Mission

BIGmama's mission is to democratize the access to bespoke AI applications, to transform companies and individuals by introducing AI in their jobs however possible.

From an original method based on the know-how developed while working with their partners for nearly a decade, they are able of identifying and extracting expertise as well as combine it with machine learning system (Hybrid AI).

What is important at BIGmama today is to be able to industrialize this methodology by proposing a software (HYKO) that automatize the extraction of expertise and its hybridization with machine learning models. In this way, this tool will allows them to build state of the art predictive applications, relatively quickly and at an affordable cost.

They believe that a large part of the future of the AI sector will be played out in the ability to articulate human expertise with the countless possibilities offered by AI models.

2.5 Vision

Data science will soon become a commodity, the arrival of large language models (which are considered foundational models in NLP) mark but a start of a new wave of foundational models. In the near future, we can expect the emergence of generalist models that are equivalent to ChatGPT but specialized in computer vision tasks, forecasting, pattern recognition, and other domains.

AI projects that take 6 months in the making, require extensive ”hyper-parameter tuning”, hundreds of experiments, task specific architectural modeling, and that come at an expensive price will soon become obsolete and outdated.

The future is going to become in the hands of those who are capable of harnessing the power of these generalist models. Algeria and the whole African continent is far behind, and its no longer a question of closing this gap step by step, a ”quantum jump” is required. BIGmama will lead this quantum jump, and will prove that excellence can beam from within the African continent.

2.6 Unique methodology

One of the valuable heritages that BIGmama acquired during the 8 years of actively developing bespoke AI applications to its clients is a the unique methodology of work at BIGmama. This methodology is centered around the following ideas :

2.6.1 AI starts with problematization.

Bespoke AI solutions development does not start with AI but with a re-framing and a ”problematization” work. Clients most often arrive with a subject (and not problems). A subject in it self is composed of many ”hidden” problems (often hidden even to the client), a good consultant at BIGmama knows that the process of ”problematization” (that of decomposing the subject into multiple problems) requires multiple iterations and attack angles, and that its success is deeply rooted to that of understanding the expertise of the clients, their needs and to the skills of the consultant to translate these problems into something addressable by the current AI models.

2.6.2 AI is a tool, not an end-goal

BIGmama firmly believes that AI should be seen as a powerful tool rather than an ultimate goal in itself. While AI technologies continues to advance rapidly, its true value lies in its ability to address and solve real-world problems and challenges faced by individuals, businesses, and society as a whole. By recognizing AI as a tool, BIGmama emphasize its role in enabling and enhancing human capabilities rather than replacing them, and emphasis that the current approaches in AI are not always the best way to solve *every* problem.

2.6.3 Hybridization

The future of AI lies in what is commonly referred to today as hybrid AI. This is a set of approaches and methodologies aiming at combining the potential of models with purely human knowledge. This hybridization allows BIGmama to put humans at the heart of technological development and to produce tools that are more efficient, easier to maintain, explain and that are far less expensive.

2.7 Software as a service (SaaS)

BIGmama is currently working on a software that intrinsically embeds its mission of democratizing AI and its unique methodology of how to do so. Hyko is an iterative software that help users formulate their problems properly using AI conversational systems powered by LLMs. It then automatically generate working applications that combines and orchestrate multiple AI models predisposed in a model-base.

Hyko can be described as a three steps iterative process :

Diagnosis and problematization (Scoping) As discussed earlier developing a bespoke AI application starts first by identifying the needs and sub-problems of your clients. Folks at BIGmama were able to identify the most important questions to ask your client in order to reach a common understanding of what can be done and how. These questions cover topics ranging from ethics and governance, success conditions, explainability of the solution, to empirical rules discovered by the experts.

Today this process of scoping is automatized with an intelligent chatbot in Hyko.

Automatic prototyping Hyko leverages the ever increasing availability of open-source AI models (Huggingface [48], Github, etc.) and the zero-shot reasoning capabilities [22] of today's LLMs to generate a first prototype solution based on the output of the scoping phase.

In order to design and execute this prototype we need first to solve the following two tasks :

1. task planning : designing how the sub problems interact and should be addressed, setting up the inputs, outputs and the dependencies of each problem (task) in the pipeline.
2. model selection : selecting the right model (available in the model base) to tackle each problem in the pipeline.

Similar to HuggingGPT [40], by solving these tasks Hyko is capable of going from a scoping of the user's problems to an executable AI prototype solution fully automatically.

Enhancement of the prototype The last step involves further customizing and enhancing the prototype to the exact needs of the client. Following this framework Building tailor-made AI solution will become cheaper faster and accessible to everyone, requiring minimal (to none) technical intervention.

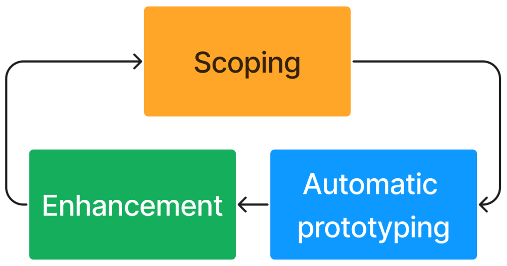

Figure 2: Iterative process of developing an AI asset in Hyko.

2.8 Knowledge representation

A very important aspect of the first step (Scoping) is the ability to represent already existing knowledge (e.g. technical documents) and make it digestible by Hyko chatbot.

This allows the chatbot to use necessary information and guide the conversation effectively.

At BIGmama we experimented with various techniques and methods for representing existing body of knowledge, first approaches involved working with ontologies and knowledge graphs :

1. Ontologies: Ontologies provide a structured way to represent knowledge by defining concepts, relationships, and properties within a specific domain. They establish a common vocabulary and enable reasoning about the knowledge. Ontologies are often represented using ontology languages like RDF (Resource Description Framework) or OWL (Web Ontology Language).

2. Knowledge graphs: Knowledge graphs organize knowledge as a network of interconnected nodes representing entities, concepts, or facts. These graphs capture not only the relationships between entities but also their attributes and contextual information. Knowledge graphs can be built using graph databases or triple stores.

However with the last advances in language modeling and information retrieval systems, document-based question answering become a more appealing option.

2.8.1 Document-based question answering

Document-based question answering (DBQA) is a machine learning task that focuses on extracting accurate and relevant answers from a given document and responding in natural language to a specific query.

It involves representing and manipulating content within textual documents, such as articles, technical documentation in order to provide precise answers to user queries.

DBQA systems typically involve a two stages process :

1. context retrieval : employing information retrieval (IR) techniques to extract relating contextual passages from the document.

2. question answering (QA) : using natural language processing (NLP) approaches to answer a query using the retrieved context.

2.9 Conclusion

In this paper we will dive deeper into the different frameworks, models and techniques used in DBQA systems (and similar QA tasks). We experimented further with different DBQA retriever-generator settings, we prepared a small benchmark for few open-source LLMs used (in combinations) with context retrieval systems.

3 State of the art

3.1 Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and human language. It involves the development of algorithms and models that enable computers to understand, interpret, and generate human language in a meaningful way.

One fundamental aspect of NLP is the representation of words as dense vectors in a high-dimensional space, known as embeddings. Word embeddings aim to capture the semantic and syntactic relationships between words, enabling algorithms to process language more effectively. These embeddings can be created using neural or statistical approaches.

Transformers, introduced by "attention is all you need" [43], are a type of deep learning model that revolutionized NLP. Transformers utilize attention mechanisms, which allow the model to attend to different parts of the input sequence when processing each element. The encoder-decoder architecture extends the transformer model to handle tasks with varying input and output sequence lengths.

As we will see NLP plays a crucial role in tackling the DBQA task, but first we will dive further into the state-of-the-art of the different embedding models and transformer architectures available in the litterateur. We will further explore the different frameworks used to solve similar tasks to DBQA.

3.1.1 Embeddings

A word embedding is a technique used in NLP to represent words as dense vectors in a high-dimensional space (an embedding). It aims to capture the semantic and syntactic relationships between words, enabling algorithms to process language more effectively.

Word embeddings can be created using two main approaches (or a combination of both): neural and statistical. Neural approaches, use neural networks to learn word representations by capturing semantic and syntactic relationships. Statistical approaches, rely on statistical features of words and their occurrences.

TF-IDF TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used weighting scheme in information retrieval and text mining tasks, designed to capture the importance of terms in a document collection. It combines two fundamental concepts: term frequency (TF) and inverse document frequency (IDF). TF refers to the number of times a term appears in a document, normalized by the total number of terms in that document. IDF, on the other hand, quantifies the rarity of a term across the entire document collection by taking the logarithm of the inverse of its document frequency.

Document frequency refers to the number of documents in the collection that contain a given term. The TF-IDF weight for a term in a document is obtained by multiplying its TF value by its IDF value. This approach assigns higher weights to terms that appear frequently in a specific document but are relatively rare across the entire collection, effectively capturing their discriminative power. the TF-IDF weight of a term t in a document d can be computed as follows:

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

where $f_{t,d}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document d, N the total number of documents in the corpus $N = |D|$ and $|\{d \in D : t \in d\}|$ is the number of documents where term t apears.

The resulting TF-IDF weights can be used for various purposes, such as document ranking, text classification, or keyword extraction, enabling the identification of important terms that characterize the content of documents in a collection.

Bert The BERT model [10] is pre-trained on large-scale corpora using unsupervised learning, followed by fine-tuning on specific downstream tasks. The resulting contextual word embeddings capture intricate semantic and syntactic relationships, enabling BERT to excel in various NLP tasks However BERT [10] uses a cross-encoder (i.e. two sentences are passed to the transformer network and the target value is predicted). This makes it unsuitable for various pair regression tasks due to too many possible combinations (finding most similar pair in a collection of n sentences requires $\frac{n*(n-1)}{2}$ inferences). Sentence-BERT (SBERT) [37], is a modification of the BERT network using siamese and triplet networks, SBERT is able to derive semantically meaningful sentence embeddings (semantically similar sentences are close in vector space).

E5 According to Huggingface embedding models leaderboard [30], the E5 [45] is the current stat-of-the-art (average performance across 62 datasets and more than 7 tasks).

MiniLM MiniLM [46] uses a simple and effective knowledge distillation method to compress large pre-trained Transformer based language models. The student model (MiniLM) is trained by deeply mimicking the teacher's self-attention [43] modules, which are the vital components of the Transformer networks.

The authors of [46] propose using the self-attention distributions and value relation of the teacher's last Transformer layer to guide the training of the student, which is effective and flexible for the student models.

The embeddings that come from this model are of high quality, cheap to compute and can be used in a number of down stream tasks.

3.1.2 Transformers

Transformers are a type of deep learning model that have revolutionized natural language processing (NLP) tasks. They were introduced by [43] and have since become the state-of-the-art approach for various NLP applications.

Attention Mechanism:

Attention mechanisms are a key component of transformers. They allow the model to focus on different parts of the input sequence when processing each element. The attention mechanism assigns weights to each element in the input sequence based on its relevance to the current element being processed.

Given an input sequence $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, the attention mechanism computes a weighted sum of all elements in $\mathbf{X}$, where the weights are determined by the relevance of each element to the current element. This can be mathematically represented as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively, and d_k represents the dimension of the key.

Multi-Head Attention:

To capture different types of information, self-attention is performed multiple times in parallel, with different learned linear projections. The outputs of these attention heads are concatenated and linearly transformed:

$$MultiHead(X) = Concat(head_1, head_2, ..., head_h)W_O$$

where $head_i = SelfAttention(XW_{Qi}, XW_{Ki}, XW_{Vi})$, W_{Qi}, W_{Ki}, and W_{Vi} are learnable weight matrices, and W_O is the output weight matrix.

Encoder Transformer:

In a transformer model, the encoder is responsible for encoding the input sequence. It consists of a stack of identical layers, each comprising a multi-head self-attention mechanism and a feed-forward neural network. The encoder processes the input sequence in parallel and captures the dependencies between different elements.

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ be the input sequence. The encoder takes $\mathbf{X}$ and produces a sequence of encoded representations $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$. The encoding process can be expressed as:

$$\mathbf{Z} = Encoder(\mathbf{X}) = EncoderLayer(\ldots(EncoderLayer(\mathbf{X})))$$

where $EncoderLayer$ represents a single layer in the encoder, and the ellipsis denotes the stacking of multiple layers.

Encoder-Decoder Architecture:

The encoder-decoder architecture extends the transformer model to tasks such as machine translation, where the input and output sequences have different lengths. The encoder processes the input sequence, while the decoder generates the output sequence.

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_m\}$ and $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$ be the input and output sequences, respectively. The encoder-decoder architecture can be summarized as:

$$\mathbf{Z} = Encoder(\mathbf{X})$$
$$\mathbf{Y}' = Decoder(\mathbf{Z}, \mathbf{Y}) = DecoderLayer(\ldots(DecoderLayer(\mathbf{Z}, \mathbf{Y})))$$

where $Decoder$ represents the decoder, $\mathbf{Z}$ is the encoded representation of $\mathbf{X}$ obtained from the encoder, and $\mathbf{Y}'$ is the predicted output sequence.

The decoder processes the output sequence $\mathbf{Y}$ and attends to the encoded representation $\mathbf{Z}$ to generate the output. The decoder also employs self-attention, allowing it to attend to previously generated elements in the output sequence.

Decoder Transformer The decoder transformer is an essential component of the transformer model for text generation tasks. It operates in an autoregressive manner, generating one token at a time. Given an encoded representation Z obtained from the encoder, the decoder attends to Z and utilizes self-attention mechanisms to capture contextual dependencies.

Writers of [33] gives a good formulation for the decoder transformer task :

Given a vocabulary $\mathcal{V}$, let $x_n \in \mathcal{V}$ for $n \in [N_{data}]$ be a dataset of sequences (imagined to be) sampled i.i.d. from some distribution P over $\mathcal{V}$. The goal is to learn an estimate $\hat{P}$ of the distribution $P(x)$. In practice, the distribution estimate is often decomposed via the chain rule as $\hat{P}(x) = \hat{P}\theta(x[1]) \cdot \hat{P}\theta(x[2], |, x[1]) \cdot \ldots \cdot \hat{P}_\theta(x[T], |, x[1 : T-1])$, where θ consists of all neural network parameters to be learned. The goal is to learn a distribution over a single token $x[t]$ given its preceding tokens $x[1 : t-1]$ as context.

By iteratively generating each token while considering the context, the decoder transformer produces high-quality and coherent text. It is widely used in various applications such as machine translation, text summarization, and question answering.

3.1.3 Similarity measures

Similarity measures are used to quantify the degree of similarity or dissimilarity between two objects or vectors. They are widely used in various domains, including information retrieval, recommendation systems, and clustering. Different similarity measures capture different aspects of similarity based on the specific requirements of the task.

Here we mention a few examples of similarity measures commonly used in the litterateur. However the selection of a suitable similarity measure is contingent upon the characteristics of the data and the particular objectives of the task being performed.

Cosine Similarity:

Cosine similarity is a commonly used similarity measure that computes the cosine of the angle between two vectors. It is particularly useful when comparing the similarity between documents, text embeddings, or high-dimensional data. The cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ can be calculated as:

$$cosine\_similarity(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where $\cdot$ denotes the dot product and $\|\cdot\|$ represents the Euclidean norm.

Euclidean Distance:

Euclidean distance is a commonly used dissimilarity measure that calculates the straight-line distance between two points in Euclidean space. It is widely used in clustering algorithms, such as k-means. The Euclidean distance between two vectors $\mathbf{A}$ and $\mathbf{B}$ of the same dimension can be computed as:

$$euclidean\_distance(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

where A_i and B_i represent the i-th elements of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively.

Jaccard Similarity:

Jaccard similarity [18] is a measure commonly used for comparing the similarity between sets. It is particularly useful in text mining and recommendation systems. The Jaccard similarity between two sets A and B is calculated as the size of their intersection divided by the size of their union:

$$jaccard\_similarity(A,B) = \frac{|A\cap B|}{|A\cup B|}$$

where $|A|$ and $|B|$ denote the cardinalities of sets A and B, respectively.

Hamming Distance:

Hamming distance [14] is a similarity measure used for comparing binary vectors of the same length. It calculates the number of positions at which the corresponding elements of two vectors are different. The Hamming distance between two binary vectors $\mathbf{A}$ and $\mathbf{B}$ can be computed as:

$$hamming\_distance(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{n}(A_i \oplus B_i)$$

where A_i and B_i represent the i-th elements of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively, and $\oplus$ denotes the bitwise XOR operation.

3.2 Many LLMs

Plethora of large language models (LLMs) are being released week upon week. Some models are made open-source : OpenLlama [13], GPT-2 [35], others are proprietary : GPT-3[4] and GPT-4 [31], some are published for commercial use such as Falcon [1], others are only for research purposes : Llama [42] and Vicuna [6].

Its hard to benchmark the performance of these models due to the wide range of applicable evaluation tasks and the difficulty of filtering out the genuine progress from grandiose claims as well as problems with current QA datasets and unrigorous evaluations [25].

Huggingface recently published an LLM leaderboard [2], where they evaluate publicly available LLMs on 4 key benchmarks :

1. AI2 Reasoning Challenge (25-shot) - a set of grade-school science questions [8].
2. HellaSwag (10-shot) - a test of commonsense inference, which is easy for humans (95%) but challenging for SOTA models [50].
3. MMLU (5-shot) - a test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more [15].
4. TruthfulQA (0-shot) - a benchmark to measure whether a language model is truthful in generating answers to questions [27].

Most of the top performing models in this benchmark are either fine-tuned Falcon [1] or fine-tuned LLama [42].

Falcon Large language models are typically trained using a combination of carefully selected high-quality data sources and filtered web data. The curation process aims to create models that perform well across a wide range of tasks. However, as the size of models increases and more training data is required, concerns arise regarding the scalability of the curation process and the availability of unique high-quality data.

The refined web dataset for Falcon llm [32] demonstrates that properly filtered and deduplicated web data alone can yield powerful language models, even outperform state-of-the-art models trained on curated datasets like The Pile [11]. The authors of this dataset[32] have released the Falcon [1] language models with 1.3/7.5 billion parameters trained on this dataset, which serve as valuable resources for this project.

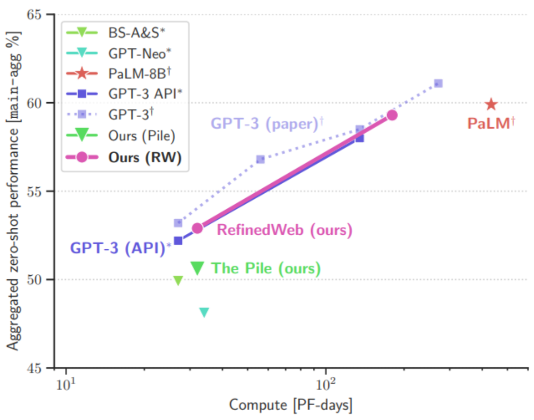

Figure 3: Models trained on REFINEDWEB [32] alone outperform models trained on curated corpora

LLaMa The authors of [42] present LLaMA, a collection of foundation language models with parameters ranging from 7 billion to 65 billion. They demonstrate that it is possible to train state-of-the-art models using publicly available datasets alone, without relying on proprietary or inaccessible data. In particular, LLaMA-13B surpasses the performance of GPT-3 (175B) [4] on most benchmarks, and LLaMA-65B is on par with other top models like Chinchilla-70B [17] and PaLM-540B [7]. By making open-sourcing the LLaMA architecture and training code and by making the weights available for researchers, plenty of other models appeared as a result of fine-tuning the original LLaMA models on various tasks and datasets, these fine-tuned models include Alpaca [41] and Vicuna [6].

Open-LLaMA [13] provides a totally open-source replication (weights and code, no research-only restriction) of the original LLaMA [42] models and which surprisingly in some cases outperforms the original models according to their benchmark [13].

Open-LLaMA is a permissively licensed open source reproduction of LLaMA [42] 7B and 13B trained on the RedPajama [9] dataset.

It was evaluated on a wide range of tasks using lm-evaluation-harness [12] (results shown in table 3.2). The LLaMA [42] results are generated by running the original LLaMA model on the same evaluation metrics. They note that the results for the LLaMA model differ slightly from the original LLaMA paper, which can be due to different evaluation protocols. Additionally, they present the results of GPT-J [44], a 6B parameter model trained on the Pile dataset [11].

The original LLaMA model was trained for 1 trillion tokens and GPT-J was trained for 500 billion tokens. from the results presented in table 3.2 OpenLLaMA exhibits comparable performance to the original LLaMA [42] and GPT-J [44] across a majority of tasks, and outperforms them in some tasks.

Task/Metric	GPT-J 6B	LLaMA 7B	LLaMA 13B	OpenLLaMA 7B	OpenLLaMA 3B	OpenLLaMA 13B
anli_r1/acc	0.32	0.35	0.35	0.33	0.33	0.33
anli_r2/acc	0.34	0.34	0.36	0.36	0.32	0.33
anli_r3/acc	0.35	0.37	0.39	0.38	0.35	0.40
arc_challenge/acc	0.34	0.39	0.44	0.37	0.34	0.41
arc_challenge/acc_norm	0.37	0.41	0.44	0.38	0.37	0.44
arc_easy/acc	0.67	0.68	0.75	0.72	0.69	0.75
arc_easy/acc_norm	0.62	0.52	0.59	0.68	0.65	0.70
boolq/acc	0.66	0.75	0.71	0.71	0.68	0.75
hellaswag/acc	0.50	0.56	0.59	0.53	0.49	0.56
hellaswag/acc_norm	0.66	0.73	0.76	0.72	0.67	0.76
openbookqa/acc	0.29	0.29	0.31	0.30	0.27	0.31
openbookqa/acc_norm	0.38	0.41	0.42	0.40	0.40	0.43
piqa/acc	0.75	0.78	0.79	0.76	0.75	0.77
piqa/acc_norm	0.76	0.78	0.79	0.77	0.76	0.79
record/em	0.88	0.91	0.92	0.89	0.88	0.91
record/f1	0.89	0.91	0.92	0.90	0.89	0.91
rte/acc	0.54	0.56	0.69	0.60	0.58	0.64
truthfulqa_mc/mc1	0.20	0.21	0.25	0.23	0.22	0.25
truthfulqa_mc/mc2	0.36	0.34	0.40	0.35	0.35	0.38
wic/acc	0.50	0.50	0.50	0.51	0.48	0.47
winogrande/acc	0.64	0.68	0.70	0.67	0.62	0.70
Average	0.52	0.55	0.57	0.55	0.53	0.57

Table 1: Evaluation of OpenLlama on the lm-evaluation-harness comparing to LLaMa and GPT-J

Vicuna [6] is an open-source chatbot trained by fine-tuning LLaMA [42] on user-shared conversations collected from ShareGPT. The preliminary evaluation conducted using GPT-4 as a judge demonstrates that Vicuna-13B achieves more than 90% quality compared to OpenAI's ChatGPT and Google Bard, surpassing models like LLaMA [42] and Stanford Alpaca [41] in over 90% of cases. The cost of training Vicuna-13B is approximately $300 (which is very impressive !). The code, weights, and an online demo of the chatbot are publicly available for non-commercial use.

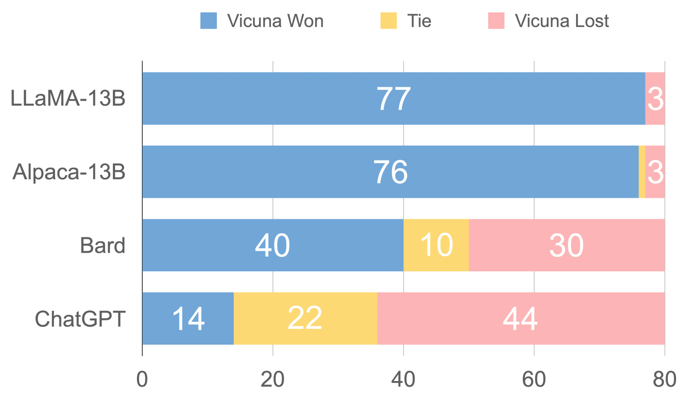

Figure 4: Response comparison of Vicuna assessed by GPT-4

3.3 Question answering

3.3.1 Open-domain QA

Open-domain Question Answering (ODQA) is a task examining the ability of models to produce answers to natural language factoid questions drawn from an open set of domains.

DrQA [5] implemented Wikipedia as its knowledge source and this choice has became a default setting for many ODQA studies since then.

However it's relevant in some cases (such as ours) to limit the knowledge source to a signal (or multiple) pre-defined documents, there is a slight difference between the two tasks, to avoid confusion we call the latter ”document-based question answering”.

Nevertheless approaches in ODQA are applicable for DBQA.

3.4 Methods for document-based question answering

3.4.1 Open-book QA

Open-book models in the field of Open-Domain Question Answering (ODQA) initially retrieve relevant documents then either extract or generate answers based on the information contained in the retrieved documents. We can distinguish mainly two approaches to tackle the ODQA problem :

1. Retriever-reader : this model works toward finding the related context in the documentation than process the retrieved context to **extract** the start/end positions of an answer. The output of the model is the selected context and the identified span of the answer in the context.

2. Retriever-generator : unlike the reader model, the generator model generate free text conditioned with the retrieved context to answer the question.

Retriever-reader Dense Passage Retrieval (DPR) [20], follows a pipeline approach, where documents are retrieved using dense embeddings. These retrieved documents are then passed to a conventional reader-re-ranker, which extracts specific spans of text as answers.

Retriever-generator Retrieval-Augmented Generation [24] is a seq2seq model that jointly learns to retrieve and generate answers. It utilizes dense retrieval and BART [23] for this purpose.

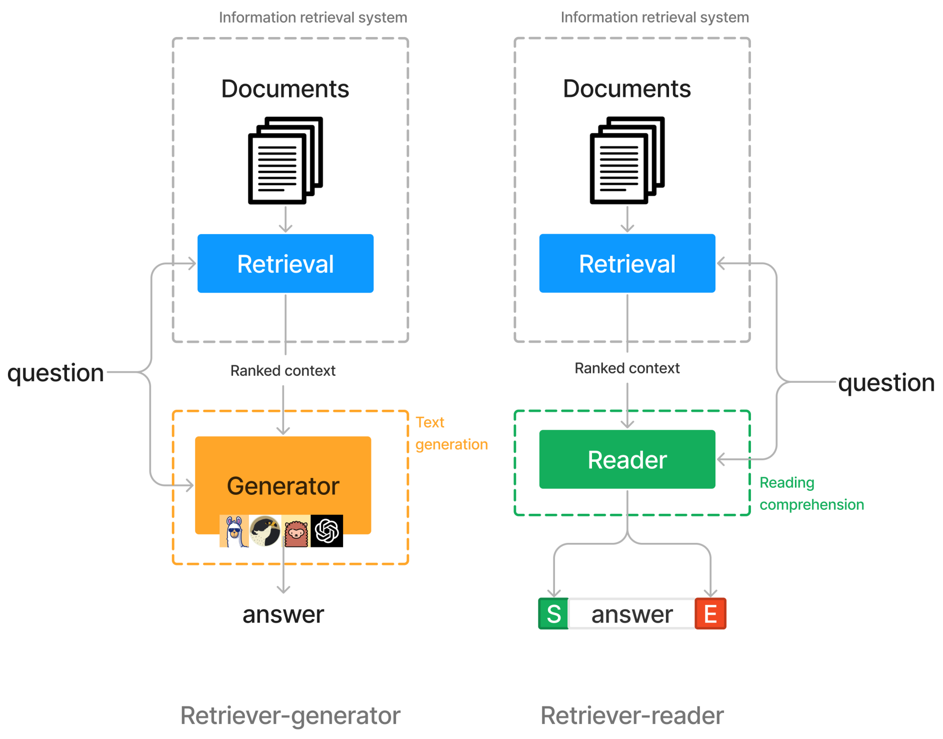

Figure 5: Overview of open-book question answering approaches.

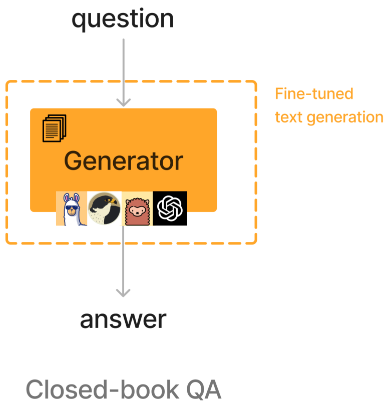

Figure 6: Overview of closed-book fine-tuned QA.

3.4.2 Closed-book QA

Large language models undergo extensive unsupervised pre-training phase using large amount of textual data. With a substantial number of parameters, these models possess the ability to memorize factual information within their weight parameters. Consequently, one is able to employ this property for question-answering tasks without relying on explicit context (shown in figure 6).

The pre-trained language models generate free-form text responses to questions, without explicitly employing reading comprehension techniques. Closed-book models can encode the given documentation within the parameters of the model itself to answer queries, rather than using a retrieval model.

Authors of [38] fine-tune a pre-trained T5 [36] model to answer questions (without access to any external context or knowledge) and were able (at the time) to compete with open-domain systems that explicitly retrieve answers from an external knowledge source when answering questions.

GPT-3 [4] has been evaluated on the closed book question answering task using the TriviaQA dataset [19] without any gradient updates or fine-tuning, the evaluation (figure 7) shows that GPT-3 match/exceed the performance of state of the art (at that time).

3.5 Retrieval models

When it comes to implementing a retriever for a retriever-generator/retriever-reader models, there is mainly two systems :

1. using classic non-learning-based TF-IDF features (“classic IR”).
2. or using dense embedding vectors of text produced by neural networks (“neural IR”).

classic IR For example DrQA [5] adopts an efficient non-learning-based search engine based on bigram hashing and TF-IDF matching.

Another approach used by BERTserini [49] consists of ranking retrieved text segments using BM25 [39], a classic TF-IDF-based retrieval scoring function.

In terms of the effect of text granularity on performance, [49] found that paragraph retrieval ¿ sentence retrieval ¿ article retrieval.

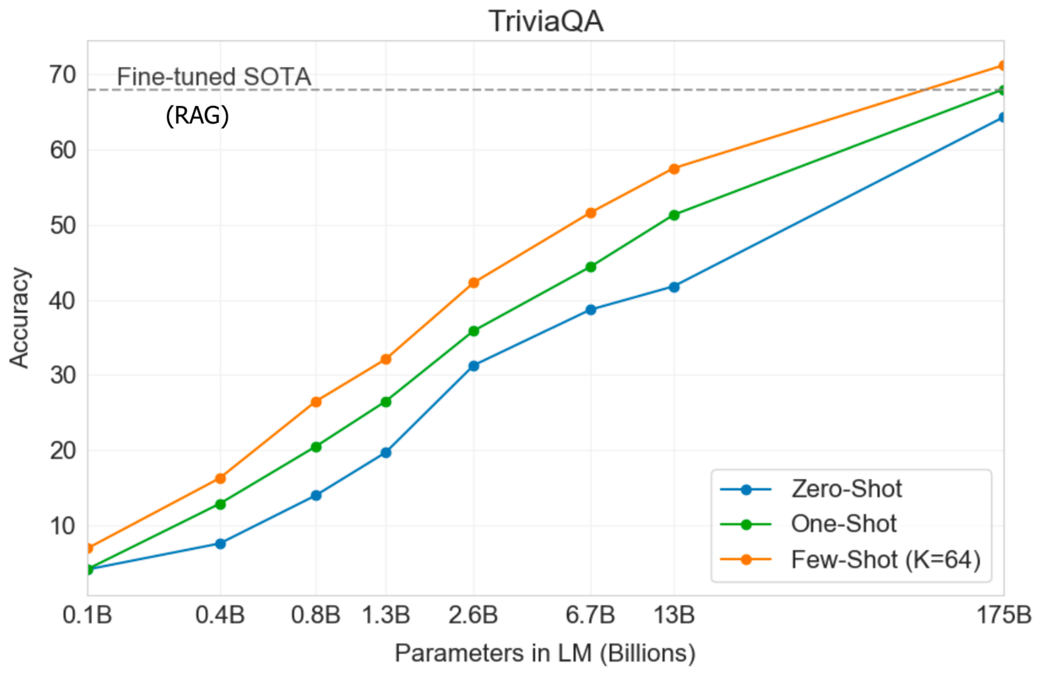

Figure 7: GPT3's performance on TriviaQA [19]. [4]

Multi-passage BERT QA model [47] uses elasticSearch with BM25 [39]. They found that splitting articles into passages with the length of 100 words by sliding window brings 4% improvements, and that splitting documents into passages without overlap causes some near-boundary evidence to lose useful contexts thus decreasing the performance.

neural IR Neural IR is a new category of methods for retrieval problems, it mainly uses dense representations of some neural network architectures (e.g. LSTM [16], BERT [10], etc).

After the arrival of many ”general language models” (to do cite llms) many IR for QA systems follow this (or a slightly different) approach :

1. Extract the dense representations of a question and a context passage by feeding them into a language model.

2. Use the dot-product of these two representations as the retrieval score to rank and select most relevant passages.

$$h_q = E_q(q)$$
$$h_z = E_z(z)$$
$$score = h_q^\top h_z$$

(however it is not necessary that neural IR out-performs classic IR [26]).

3.6 Reader Models

The reader model objective is to learn solve reading comprehension tasks, extract an answer for a question from a retrieved context. Here we only discuss approaches for machine comprehension using neural networks since it yields best performing results.

4 Implementation and results

4.1 Introduction

In our Implementation we chose a retriever-generator approach, we experimented with different combinations of embedding and Language models, as well as different similarity measures and text splitting strategies.

We evaluated the different systems using a sample from the SQuAD [34] dataset while using GPT-3 [4] as a scoring model (more about that in the sections below).

Similar work LlamaIndex [28] tested out few indexing methods (embeddings, LLMs and retrieval systems) following a similar framework to ours. However their focus was on OpenAI proprietary language models (GPT-3 and GPT-4) [4, 31]. Our focus in this project on the other hand is on using open-source embedding models and LLMs that can fully run on-premise, since model and data sovereignty are a critical aspect for our use-case.

In the following sections we will provide more details on our implementation as well as the evaluation results we got.

4.2 challenges

The following challenges arise when using LLMs for DBQA :

1. Small context window: LLMs typically have a limited context window, which means they can only consider a limited number of preceding tokens when generating a response. In DBQA, where the answer may depend on information scattered throughout a document, this limited context window can make it difficult for LLMs to capture the necessary context and provide accurate answers. Relevant information may fall outside the context window, leading to incomplete or incorrect responses. This is why we need information retrieval systems when choosing the context.

2. Hallucination: LLMs can sometimes generate responses that are plausible-sounding but factually incorrect. This phenomenon, known as hallucination, can be a significant issue in DBQA. Since LLMs rely on statistical patterns and language modeling rather than true understanding, they may generate answers that sound reasonable but are not grounded in the actual content of the document. Hallucination can mislead users and compromise the reliability of the DBQA system.

3. Storing the models and computational demand for inference: LLMs, especially large-scale models, require substantial computational resources for both training and inference. Storing these models and performing inference can be challenging due to their size and computational demands. These demands can pose practical challenges for deploying LLM-based DBQA systems at scale.

We were able to workaround the first challenge using IR system based on neural-embedding. For the second challenge we tried using small temperature (a hyperparameter used to control the randomness of the generated text) and presence penalty values (high presence penalty values results in the model being more likely to generate tokens that have not yet been included in the generated text and vice-versa). Using smaller models (7B parameters each) required around 36GB of memory, we ran the inferences on a multi-GPU machine.

4.3 Description

The systems that we experimented with follows the retriever-generator framework which consists mainly of two parts :

A retrieval system that takes as input a document (text) and a query from the user, splits the document into sentences. Each sentence from the document along the query of the user are going to be embedded by one of the chosen embedding models. We measure the one-to-one similarity (using cosine similarity) between the query embedding and each sentence embedding and we retrieve the top K sentences to be used a context.

A conditioned generative system it takes as input the retrieved contexts and the query to construct a prompt that condition the generation using one of the chosen open-source LLMs. In table 2 we show that we tested all the possible combinations of the different embedding models and LMs at hand.

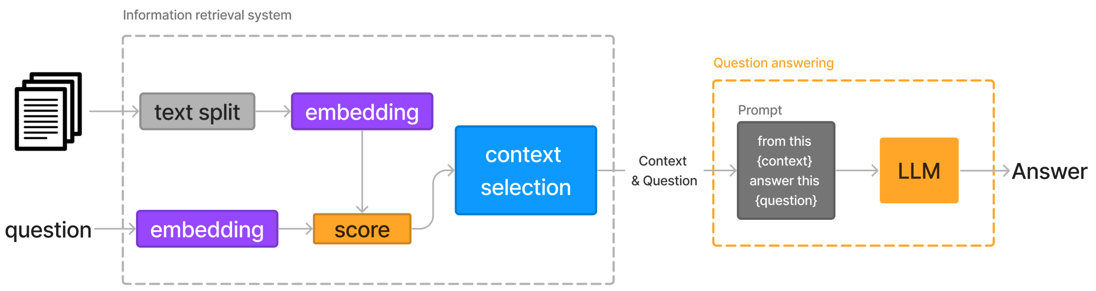

Figure 8: Highlevel overview of the whole system.

4.4 Evaluation

We bench-marked the different systems on the SQuAD dataset [34]. We sampled 20 examples from the dataset with their contexts and answers (examples can be found in annex 6.2), for each question and context we tested a different embedding model and LLM combination, each response is then evaluated using GPT-3 [4] (a score of +1 is given if the answer is correct 0 otherwise), results are shown in table 2.

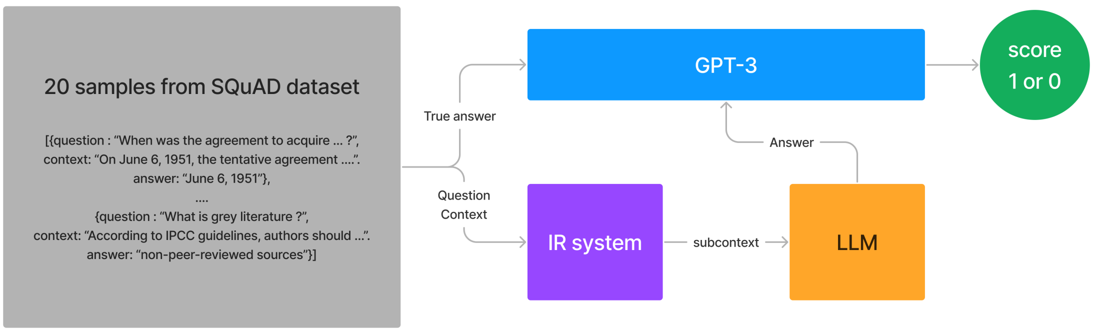
20 samples from SQuAD dataset
[{question : "When was the agreement to acquire ... ?",
context: "On June 6, 1951, the tentative agreement ....".
answer: "June 6, 1951"},
....
{question : "What is grey literature ?",
context: "According to IPCC guidelines, authors should ...".
answer: "non-peer-reviewed sources"}]
True answer
Question Context
GPT-3
score 1 or 0
Answer
IR system
subcontext
LLM

Figure 9: Evaluation schema with GPT-3 [4].

	Language model		
Embedder	**Vicuna**	**Falcon**	**OpenLlama**
Bert	**0.4**	0.25	0.35
MiniLM	0.35	0.25	0.3
E5	0.35	0.25	0.3

Table 2: Benchmark : average score of each system when tested on 20 samples from SQuAD [34] dataset, and evaluated using GPT-3 [4].

4.5 Discussion

4.6 Implementation details

In this section, we provide general details about the implementation of our benchmarking system. We also describe the specific setup used for our experiments.

1. **Python version:** We implemented the system using Python 3.10.11, which provided us with the necessary language features and libraries for our project.
2. **Operating system:** The implementation was developed and tested on Ubuntu 20.04.6 LTS x86_64.
3. **Graphics Processing Units (GPUs):** To leverage the power of parallel processing, our setup included a combination of GPUs. We utilized one NVIDIA GeForce GTX 1080 Ti and two NVIDIA GeForce RTX 3060 GPUs for faster inference.
4. **Memory :** The system had access to a substantial amount of RAM, specifically 126GB. This allowed us to load the weights of the different LLMs easily.

IR system details

The Information Retrieval (IR) system employed a two-step process for document retrieval:

1. **Splitting:** We tested different text splitting mechanisms, on the token level with overlap using the tiktoken package, on the sentence level using a metaheuristic developed by the curators of Europarl [21] (pseudo code available on annex 6.1.6) and on the paragraph level by concatenating multiple sentences (while making sure that the last sentence ends with a dot).
2. **Embedding:** We used a BERT [10] and E5 [45] from HuggingFace [48] (bert-base-cased-squad2 and e5-small respectively) and a MiniLM [46] model from the SentenceTransformer package [37].

During the evaluation phase we fix the text splitting to sentence level and the similarity measure to cosine similarity and $k = 1$ (the number of retrieved context sentences), to make the evaluation a little more rigorous.

Language models details

We employed a batching strategy processing multiple input sequences simultaneously, reducing the overall processing time.

Using the OpenLlama project [13] weights, we employed the LlamaForCausalLM model and the LlamaTokenizer from HuggingFace [48] to load and run the model. During generation, a maximum of 30 new tokens were allowed, and the generation temperature was set to 0.0 for deterministic output.

For Falcon [1] we employed the AutoTokenizer and the ”text-generation” pipeline from the HuggingFace [48]. During generation, we used a temperature of $3e - 4$. The maximum length of generated tokens was set to 30.

For Vicuna [6] we used FastChat [29] implementation. We use their implementation for chat input and output as a workaround for batched inference (shown in annex 6.1.2). The maximum number of new tokens and temperature are 30 and 0.0 respectively

Evaluation details

The evaluation of our benchmark involved sampling 20 random samples from the SQuAD [34] dataset, and using GPT-3 [4] as a scoring model. We use LMQL [3] programming framework to constrain the evaluation output of the scoring model (shown in annex 6.1.4). LMQL uses token masking to prevent the generation of tokens that violate the specified constraints. If a constraint cannot be satisfied, the corresponding branch of generation is pruned, and the model continues generating in the remaining valid branches. This helps improve efficiency and constrain the model to output only necessary tokens (in our case it’s 1 or 0).

User interface

- backend - frontend

5 Conclusion

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40b: an open large language model with state-of-the-art performance. 2023.

[2] Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard, 2023.

[3] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. *PLDI '23*, 2022.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions, 2017.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[8] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[9] Together Computer. Redpajama-data: An open source recipe to reproduce llama training dataset, April 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.

[12] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster,

Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021.

[13] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023.

[14] Richard W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[18] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[19] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017.

[20] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[21] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation, 2005.

[22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[23] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[25] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. Question and answer test-train overlap in open-domain question answering datasets, 2020.

[26] Jimmy Lin. The neural hype and comparisons against weak baselines, 2018.

[27] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[28] Jerry Liu. Llamaindex, 11 2022.

[29] lm sys. Fastchat. https://github.com/lm-sys/FastChat, 2023. Online; accessed 19 June 2023.

[30] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[31] OpenAI. Gpt-4 technical report, 2023.

[32] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.

[33] Mary Phuong and Marcus Hutter. Formal algorithms for transformers, 2022.

[34] purkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[37] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[38] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model?, 2020.

[39] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Bruce W. Croft and C. J. van Rijsbergen, editors, *SIGIR '94*, pages 232–241, London, 1994. Springer London.

[40] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023.

[41] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[44] Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, May 2021.

[45] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

[46] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[47] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering, 2019.

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[49] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, 2019.

[50] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

6 Annexes

6.1 Code

6.1.1 initializing the different LMs

```python
from abc import ABC, abstractmethod

import gc

import torch
import transformers
from transformers import (
    AutoModelForCausalLM,
    AutoTokenizer,
    LlamaForCausalLM,
    LlamaTokenizer
)


class LM(ABC):
    @abstractmethod
    def __call__(self, requests: list[str], **kwargs: any) -> list[str]:
        pass


class Llama(LM):
    def __init__(self, model_path: str = "models/openllama/7B") -> None:
        super().__init__()
        self.model_path = model_path
        self.tokenizer = LlamaTokenizer.from_pretrained(model_path)
        self.model = LlamaForCausalLM.from_pretrained(
            model_path,
            torch_dtype=torch.float16,
            device_map="auto",
        )

    def __call__(self, requests: list[str]) -> list[str]:
        outputs = []
        for prompt in requests:
            input_ids = self.tokenizer(prompt, return_tensors="pt").input_ids
            generation_output = self.model.generate(
                input_ids=input_ids,
                max_new_tokens=30,
                temperature=0.0,
            )
            output = self.tokenizer.decode(
                generation_output[0], skip_special_tokens=True
            )
            output = output.replace(prompt, "")  # eq : return_full_sequence=False
            outputs.append(output)
            print(output)
        return outputs


class Falcon(LM):
    def __init__(self, model_path: str = "models/falcon/7B/snapshots/falcon") -> None
    :
        super().__init__()
        gc.collect()
        torch.cuda.empty_cache()
        self.model_name = model_path

    def __call__(self, requests: list[str]) -> list[str]:
        tokenizer = AutoTokenizer.from_pretrained(self.model_name)
        pipeline_ = transformers.pipeline(
            "text-generation",
            model=self.model_name,
            tokenizer=tokenizer,
            torch_dtype=torch.bfloat16,
            trust_remote_code=True,
            device_map="auto",
        )
        sequences = pipeline_(
            requests,
            max_new_tokens=30,
            do_sample=True,
            top_k=10,
            temperature=3e-4,
            num_return_sequences=1,
            eos_token_id=tokenizer.eos_token_id,
            return_full_text=False,
        )
        outputs = [seq[0]["generated_text"] for seq in sequences]
        return outputs
```

6.1.2 vicuna workaround

```
import torch
from fastchat.model.chatglm_model import chatglm_generate_stream
from fastchat.model.model_adapter import get_conversation_template, load_model
from fastchat.serve.inference import generate_stream

from src.lm import LM


class SimpleChatIO:
    """this is a workaround to use fastchat for batched inference"""

    def __init__(self, requests: list[str]) -> None:
        self.requests = requests

    def prompt_for_input(self) -> str:
        return self.requests.pop(0) if self.requests else ""

    def stream_output(self, output_stream):
        pre = 0
        for outputs in output_stream:
            output_text = outputs["text"]
            output_text = output_text.strip().split(" ")
            now = len(output_text) - 1
            if now > pre:
                print(" ".join(output_text[pre:now]), end=" ", flush=True)
                pre = now
        print(" ".join(output_text[pre:]), flush=True)
        return " ".join(output_text)


def chat_loop(
    model_path: str,
    device: str,
    num_gpus: int,
    max_gpu_memory: str,
    load_8bit: bool,
    cpu_offloading: bool,
    temperature: float,
    repetition_penalty: float,
    max_new_tokens: int,
    chatio: SimpleChatIO,
    debug: bool,
):
    # Model
    model, tokenizer = load_model(
        model_path, device, num_gpus, max_gpu_memory, load_8bit, cpu_offloading,
   debug
    )
    is_chatglm = "chatglm" in str(type(model)).lower()
    is_fastchat_t5 = "t5" in str(type(model)).lower()

    # Hardcode T5 repetition penalty to be 1.2
    if is_fastchat_t5 and repetition_penalty == 1.0:
        repetition_penalty = 1.2

    while True:
        conv = get_conversation_template(model_path)  # reset conversation

        try:
            inp = chatio.prompt_for_input()
        except EOFError:
            inp = ""
        if not inp:
            print("exit...")
            break

        conv.append_message(conv.roles[0], inp)
        conv.append_message(conv.roles[1], None)

        if is_chatglm:
            generate_stream_func = chatglm_generate_stream
            prompt = conv.messages[conv.offset :]
        else:
            generate_stream_func = generate_stream
            prompt = conv.get_prompt()

        gen_params = {
            "model": model_path,
            "prompt": prompt,
            "temperature": temperature,
            "repetition_penalty": repetition_penalty,
            "max_new_tokens": max_new_tokens,
            "stop": conv.stop_str,
            "stop_token_ids": conv.stop_token_ids,
```

```python
            "echo": False,
        }

        output_stream = generate_stream_func(model, tokenizer, gen_params, device)
        outputs = chatio.stream_output(output_stream)

        yield outputs

        if debug:
            print("\n", {"prompt": prompt, "outputs": outputs}, "\n")


class Vicuna(LM):
    def __init__(self, model_path: str = "models/vicuna/7B") -> None:
        super().__init__()
        self.model_path = model_path
        self.device = "cuda" if torch.cuda.is_available() else "cpu"
        self.num_gpus = 3
        self.max_gpu_memory = None
        self.load_8bit = False
        self.cpu_offloading = False
        self.max_new_tokens = 20

    def __call__(
        self,
        requests: list[str],
    ) -> list[str]:
        chatio = SimpleChatIO(requests=requests)
        outputs = [
            output
            for output in chat_loop(
                self.model_path,
                self.device,
                num_gpus=self.num_gpus,
                load_8bit=self.load_8bit,
                temperature=0.0,
                repetition_penalty=1.0,
                max_new_tokens=20,
                debug=False,
                chatio=chatio,
                cpu_offloading=self.cpu_offloading,
                conv_template=None,
                max_gpu_memory=None,
            )
        ]
        return outputs
```

6.1.3 initializing the different embedding models

```python
from abc import ABC, abstractmethod

import torch
import torch.nn.functional as F
from sentence_transformers import SentenceTransformer
from torch import Tensor
from transformers import AutoModel, AutoTokenizer, BertForQuestionAnswering

from src.similarity_measure import CosineSimilarity


class Embedder(ABC):
    @abstractmethod
    def embed(self, texts: list[str]) -> list[list[float]]:
        """Embed a list of contexts"""
        pass

    def embed_query(self, texts: list[str]) -> list[list[float]]:
        """Embed a list of queries"""
        return self.embed(texts)

    def embed_context(self, texts: list[str]) -> list[list[float]]:
        """Embed a list of contexts"""
        return self.embed(texts)


class MiniLM(Embedder):
    def __init__(self) -> None:
        super().__init__()
        self.model = SentenceTransformer("all-MiniLM-L6-v2", device="cpu")

    def embed(self, texts: list[str]) -> list[list[float]]:
        embeddings = self.model.encode(texts)
        return embeddings.tolist()


class Bert(Embedder):
    def __init__(self) -> None:
        super().__init__()
        self.tokenizer = AutoTokenizer.from_pretrained("deepset/bert-base-cased-squad2")
        self.model = BertForQuestionAnswering.from_pretrained(
            "deepset/bert-base-cased-squad2"
        )

    def embed(self, texts: list[str]) -> list[list[float]]:
        inputs = [self.tokenizer(text, return_tensors="pt") for text in texts]
        with torch.no_grad():
            outputs = [
                self.model(**input, output_hidden_states=True) for input in inputs
            ]
            # take the average of the last hidden-state of each token to represent the sentence
            outputs = [
                output.hidden_states[-1].mean(dim=1).flatten().tolist()
                for output in outputs
            ]

        return outputs


class E5(Embedder):
    def __init__(self) -> None:
        super().__init__()
        self.tokenizer = AutoTokenizer.from_pretrained("intfloat/e5-small")
        self.model = AutoModel.from_pretrained("intfloat/e5-small")

    def average_pool(
        self, last_hidden_states: Tensor, attention_mask: Tensor
    ) -> Tensor:
        last_hidden = last_hidden_states.masked_fill(
            ~attention_mask[..., None].bool(), 0.0
        )
        return last_hidden.sum(dim=1) / attention_mask.sum(dim=1)[..., None]

    def embed(self, texts: list[str]) -> list[list[float]]:
        batch_dict = self.tokenizer(texts, padding=True, return_tensors="pt")

        outputs = self.model(**batch_dict)
        embeddings = self.average_pool(
            outputs.last_hidden_state, batch_dict["attention_mask"]
        )

        # (Optionally) normalize embeddings
        embeddings = F.normalize(embeddings, p=2, dim=1)
        return embeddings.detach().tolist()

    def embed_query(self, texts: list[str]) -> list[list[float]]:
        return self.embed(["query :" + text for text in texts])

    def embed_context(self, texts: list[str]) -> list[list[float]]:
        return self.embed(["passage :" + text for text in texts])
```

6.1.4 Evaluation using GPT-3 [4] with LMQL framework [3]

```python
import os

import dotenv
import lmql
import openai

dotenv.load_dotenv()
openai.api_key = os.getenv("OPENAI_API_KEY")


@lmql.query
async def evaluate_gpt3(prediction: str, ground_truth: str):
    '''argmax
        """output 1 if the student answer is similar to true answer, 0 otherwise
        ignore small differences
        student answer: {prediction}
        true answer: {ground_truth}
        evaluation : [EVALUATION]"""
    from
        "openai/text-davinci-003"
    distribution
        EVALUATION in ["1", "0"]
    '''
```

6.1.5 Index construction

```python
from src.embedding import Embedder
from src.text_splitter import Splitter
from src.utils.utils import dump_pickle, read_pickle


class Index:
    def __init__(
        self,
        embedder: Embedder,
        splitter: Splitter,
    ) -> None:
        super().__init__()

        self.splitter = splitter
        self.embedder = embedder

    def __call__(self, document: str, index_path: str | None = None) -> list[str]:
        """returns an indexed document"""
        if document:
            self.index = index = self.index_document(document)
        elif index_path:
            self.index = index = self.load_index(index_path)
        else:
            raise ValueError("Either document or an index_path must be provided")

        return index

    def index_document(self, document: str) -> list[tuple]:
        chunks = self.splitter.split(document)
        embeddings = self.embedder.embed_context(chunks)
        return list(zip(chunks, embeddings))

    def save_index(self, path: str) -> None:
        assert self.index, "Index is empty"
        dump_pickle(self.index, path=path)

    def load_index(self, path: str) -> list[tuple]:
        return read_pickle(path=path)
```

6.1.6 Sentence splitting heuristic [21] pseudocode

```
    function splitIntoSentences(text):
    sentences = []
    words = splitTextIntoWords(text)
    currentSentence = ""

    for i = 0 to length(words) - 1:
        word = words[i]
        nextWord = words[i + 1] if i + 1 < length(words) else ""

        if isEndOfSentence(word, nextWord):
            currentSentence += word
            sentences.append(currentSentence)
            currentSentence = ""
        else:
            currentSentence += word + " "

    return sentences

function isEndOfSentence(word, nextWord):
    if word ends with period, question mark, or exclamation mark:
        if nextWord starts with uppercase letter or sentence starter punctuation:
            return true

    if word ends with multiple consecutive dots:
        if nextWord starts with uppercase letter or sentence starter punctuation:
            return true

    if word ends with punctuation inside quotes or parentheses:
        if nextWord starts with possible sentence starter punctuation and uppercase
   letter:
            return true

    if isHonorific(word) and nextWord is empty:
        return false

    if isUppercaseAcronym(word):
        return false

    if word ends with period:
        if nextWord starts with uppercase letter or sentence starter punctuation:
            return true

    return false

function isHonorific(word):
    if word is a known honorific:
        return true
    return false

function isUppercaseAcronym(word):
    if word consists of uppercase letters and ends with period:
        return true
    return false

function splitTextIntoWords(text):
    words = []
    currentWord = ""

    for character in text:
        if character is whitespace:
            if currentWord is not empty:
                words.append(currentWord)
                currentWord = ""
        else:
            currentWord += character

    if currentWord is not empty:
        words.append(currentWord)

    return words
```

6.2 Experiments

Results sample