

# Deep Learning Project

## 2024-2025

### Table of Contents:

<b>Introduction</b>	2
<b>I. Literature Review</b>	2
1. Historical Perspective on LLMs	2
2. Perspectives and relationships with this paper	3
<b>II. Dataset analyses and accuracy methodology approach</b>	3
1. Dataset analysis	3
2. Accuracy methodology approach	3
<b>III. Models and tests methodology</b>	4
1. The models used	4
i. Direct answers models	4
ii. Two steps models	4
iii. Hybrid model	5
2. Test methodology	5
i. Methodology for direct answers' models	5
a. Only rules	5
b. Rules – examples	6
ii. Methodology for two steps models	6
iii. Methodology for hybrid model	6
<b>IV. Results, observations and synthesis</b>	7
1. Final results	7
i. Direct answer models	7
a. Results	7
b. Observations	7
ii. Two steps models	8
a. Results	8
b. Observations	8
iii. Hybrid model	8
a. Results	8
b. Observations	9
2. Synthesis	10
i. Error types	10
ii. Performance Gap Between Local and Cloud-Based Models	11
iii. The Retrieval Paradox: More Information Doesn't Always Mean Better Answers	11
<b>V. Going beyond</b>	12
1. Improvements and limits	12
i. Limits	12
ii. Improvements	12
2. Ethical considerations	12
<b>Bibliography</b>	13
<b>Appendixes</b>	14

## **Introduction**

Since the introduction of the Transformers architecture in 2017, the way we understand Natural Language Processing has shifted. The GPT models by OpenAI and Llama by Meta are good at understanding and generating texts with a focus on different tasks, Llama is meant to deal with longer texts and has a more encyclopedic purpose.

In this project, we analyze the “exact match” performance of these models for the task of question answering, that is interpreting a question in natural language, searching information in a database and generating answers.

One major challenge in question answering is that models generate answers based on their internal knowledge, but we have no control over what they know. Some models can also hallucinate wrong information. To address this, we compare different ways to retrieve external knowledge and combine it with the models to improve factual accuracy.

We split the methods used to compare the performance on the task in three categories: the direct answers models where we executed prompts directly through the LLM, the two-steps models that relied on the large database from Pyserini to extract a large context before formatting the responses using Roberta and finally, hybrid models consisting of the pure information-retrieval of the two-steps models before feeding the data in the LLMs with a specific prompt. This distinction is important because LLMs have no control over their sources of information making retrieval-based methods a potential way to improve factual accuracy.

Our experiments show that GPT-3.5 significantly outperforms the other models, especially with well-tuned prompts. LLaMA 3.2 performs worse and does not benefit much from prompt optimization. BM25-based approaches provide an alternative but remain limited in effectiveness.

## **I. Literature Review**

### **1. Historical Perspective on LLMs**

The Transformer architecture, introduced in 2017 with the paper "Attention Is All You Need" (Vaswani et al., 2017), marked a paradigm shift in Natural Language Processing. Prior to this, recurrent neural network (RNN) models dominated the field and data was processed sequentially. The Transformer architecture, however, made parallelization possible, significantly improving efficiency and performance. This advancement was made possible by the progress in GPU technology at the same time.

All three large language models (LLMs) discussed in this project - GPT-2, LLaMA 3.2, and GPT-3.5 - are decoder-only models based on the Transformer architecture, albeit with specific modifications and enhancements. While the datasets used to train these models have not been fully disclosed, GPT-2 and LLaMA 3.2 have made their model weights publicly available. This was no longer the case with GPT-3.5, but recent news are leading the way for more open source LLM research.

The release of GPT-2 (Radford et al., 2019) demonstrated the power of unsupervised fine-tuning on large datasets. Trained on the corpus WebText, GPT-2 outperformed supervised learning models on specific tasks GPT-2 wasn't trained for. With 1.5 billion parameters, GPT-2 sets the first steppingstones of scalable machine learning.

Four years after GPT-2, Meta introduced a 70-billion-parameter fine-tuned model, LLaMA 3.2 (Touvron et al., 2023). This model represented an exponential increase in the scale of data and computational resources used for training LLMs. LLaMA 3.2 was optimized for complex tasks and supported a context window four times larger than GPT-2. It was optimized for understanding complicated texts.

The GPT-3.5 and GPT-4 models (OpenAI, 2023) represented the latest evolution in LLMs at the time, incorporating reinforcement learning from human feedback (RLHF) for fine-tuning. The new performance benchmarks these models set came with a new milestone for scaling of data and resources required, with 200 billion to 1 trillion parameters.

As LLMs have grown in scale and complexity, their applications have also expanded, particularly in question answering. While some models rely solely on their internal knowledge, others integrate retrieval-based methods to improve accuracy and factual consistency. This distinction is central to our evaluation, as we compare different strategies to assess their exact match performance.

## 2. Perspectives and relationships with this paper

The development of LLMs has introduced different approaches to question answering, with some models focusing on direct text generation while others integrate retrieval-based methods for more structured responses. In this project, we evaluate these models based on their exact match performance, testing direct answer generation, retrieval-augmented approaches, and hybrid methods combining retrieval with model-generated responses.

## II. Dataset analyses and accuracy methodology approach

### 1. Dataset analysis

As previously explained, our objective is to test several LLM models to predict precise and concise answers from a fixed set of answers. The dataset we are working on is the NQ-Open dataset (3610 questions with associated answers). Because of its forms, the dataset brings several difficulties because of its structure. Indeed, several types of answers can be expected. We identified four types: unique or several answers -one token- (1), unique or several answers -several token- (2), unique or several answers with special signs -one token- (3), unique or several answers with special signs -several token- (4). We also have combinations of all of them (5).

Type	Question	Answer
1.1:	"What is the meaning of the name Comanche"	['enemy']
1.2:	"Who had a baby at 100 in the bible"	['Sarah', 'Abraham']
2.1:	"Who played booster in jingle all the way"	['Curtis Armstrong']
2.2:	"Where did the river thames start and end"	['Thames Estuary, North Sea']
3.1:	"When did houston go to the american league"	['2013']
3.2:	"When was the minimum wage established in the united states"	['1938', '1933', '1912']
4.1:	"When does body temperature tend to be lowest"	['11 p.m. to 3 a.m.']
4.2:	"Wow much money did the film titanic make"	['\$2.18 billion', '\$2.187 billion']
5.1:	"When is the last time the philadelphia won the superbowl"	['Super Bowl LII', '2017']
5.2:	"What is the current mac os operating system"	['10.13.4', 'macOS High Sierra']

Because of this variety of answer shapes, a clear evaluation methodology is required.

### 2. Accuracy methodology approach

We are asked to evaluate our results through an "exact match" methodology. Because of the diversity of expected answers, a perfect match, in the sense of the exact same words in the same order, does not make a lot of sense, particularly in the case of several possibles answers. We therefore decided to use an exact match function checking if the answer returned corresponds to one of the elements within the list<sup>1</sup>. However, this match must be exact in terms of capital letters, punctuation and/or specific characters. We assign a score of 1 if the returned answer matches exactly one of the possible answers, and 0 otherwise. A score of 39% indicates that 39% of the answers met this criterion and the remaining 61% did not.

We define our accuracy with the formula:

$$Accuracy = \frac{\#\{exact\ match\ returns\ 1\}}{\#\{questions\ asked\}}$$

<sup>1</sup> See details in Appendix 1

### III. Models and tests methodology

#### 1. The models used

We tested only pretrained models. We then distinguish three types of approaches. First, we tested direct answer models by only influencing them through a prompt, called “direct answers models” (i). We then tested two different retrieval methods and dealt with the shape of the outputs. These models will be called “two-steps models” (ii). We then did a combination of both methodologies i.e. using a retrieval method before using a direct answer model with a specific prompt and formatting the response, called “hybrid models” (iii).

##### i. Direct answers models

We respectively tested GPT2, LLaMA 3.2 and GPT3.5. Two main criteria guided our choices: we wanted to compare older LLM model to more recent ones, as well as local LLM models.

**GPT2 (local):** As this is not a recent model, it was important for us to test “ancient” model to compare them with the more recent ones we considered in order to address the improvement in the models.

**LLaMA 3.2 (local):** A more recent open-source model from Meta. We wanted to evaluate its sensitivity to prompts and its ability to structure answers correctly.

**GPT3.5 (API):** We chose it to evaluate the quality of our methodologies, GPT3.5 being a very efficient model.

However, these models have no control over their data sources which makes them unreliable for retrieving up-to-date or fact-checked answers. To address this, we explore retrieval-based methods, which allows us to supply the model with external, reliable sources before generating a response.

##### ii. Two steps models

We used the database from Pyserini. We used two retrieval models: BM25 and BM25 combined with an all-MiniLM-L6-v2. We then formatted the response using Roberta.

**BM25:** This model ranks documents based on the frequency and distribution of query terms. It assumes that relevant documents contain exact keyword matches, weighting term frequency and document length to estimate relevance. BM25 serves as the baseline retrieval method in this study. Once a query is issued, BM25 retrieves the top-k documents maximizing its ranking score which provides a quick initial rank before refining through semantic re-ranking or generative models. This approach is computationally efficient and widely used in information retrieval due to its robustness to handle keyword-based search.

**all-MiniLM-L6-v2:** This model improves over BM25 by capturing semantic meaning rather than relying solely on keyword matching. Instead of ranking documents based on words frequency, it encodes both queries and documents into dense vector representations and measures their similarity using cosine similarity. This allows it to retrieve relevant results even when different words are used to express the same idea.

However, computing embeddings and cosine similarities for all documents is too slow on a personal machine. To make it possible, we first use BM25 for an initial ranking to quickly select the top-k documents. Then, we apply all-MiniLM-L6-v2 to re-rank these candidates based on their semantic relevance, improving the quality of retrieved answers while keeping computation time reasonable.

**Roberta:** This is a transformer-based question-answering model trained to extract precise answers from a given context. We picked this model because unlike generative models, Roberta ensures concise and structured responses, alignment with dataset annotation style as well as minimizing nonsense.

### iii. Hybrid model

We did two types of retrieval: BM25 paired with the best document ( $k=1$ ), BM25 paired with embedding-selection of the best three documents. We then give them to GPT2, LLaMA 3.2 and GPT3.5 with a specific prompt and formatted the answer through Roberta.

## 2. Test methodology

We first selected randomly five sets of questions from our dataset. As we wanted the union of these sets to cover 50% of the questions of the dataset, we randomly selected five set of unique 361 questions in the dataset. These sets were saved and all the results in this paper is the average success and execution time of each approach on each set of 361 questions.

### i. Methodology for direct answers' models

The main variable to test and assess the quality of these models lies in the prompt used. We therefore tested them with four prompts: two strict rule-based instructions "only rules" (a) and two rule-based instructions combined with examples "rules-examples" (b), each time on the five set of 361 questions as was done previously.

#### a. *Only rules*

Models were tested with the following two prompts. The goal is to understand whether precise guidance leads to improvements.

Prompt	
Only Rules 1	Only Rules 2
<p>You are a Q&amp;A system.</p> <p>When you reply, go straight to the point.</p> <p>Do NOT even make sentences, just give ONE word, the relevant one (or the name, date etc if necessary)</p> <p>Do NOT explain anything.</p> <p>Now answer the following question:</p> <p>Question: {question} Answer:</p>	<p>You are a concise and precise Q&amp;A system. Your goal is to return <b>accurate</b> answers in <b>three words or less</b>.</p> <p>### <b>Rules</b>:</p> <ol style="list-style-type: none"> <li><b>Capitalization</b>: <ul style="list-style-type: none"> <li>- Capitalize only if the term is a proper noun, the name of someone, a place, or follows Wikipedia capitalization rules.</li> </ul> </li> <li><b>Punctuation</b>: <ul style="list-style-type: none"> <li>- Do NOT use punctuation or unnecessary characters.</li> <li>- NEVER use a period (.) or any punctuation at the end of your answer.</li> <li>- If unsure about punctuation, return the most probable format based on the dataset.</li> </ul> </li> <li><b>Answer format</b>: <ul style="list-style-type: none"> <li>- The answer must be <b>precise, factual, and certain</b>.</li> <li>- If multiple valid answers exist, return only the <b>first</b> valid one.</li> <li>- Do NOT explain or justify your response. Output only the answer.</li> </ul> </li> <li><b>Numbers &amp; Dates</b>: <ul style="list-style-type: none"> <li>- Return numbers in <b>digit format</b> (e.g., "5" instead of "five").</li> <li>- <b>Dates must be formatted as</b>: <ul style="list-style-type: none"> <li>- "YYYY" (if only year is known).</li> <li>- "DD Month YYYY" (if full date is needed).</li> </ul> </li> <li>- Use <b>English month names</b> (e.g., "14 December 1972" instead of "14 Décembre 1972").</li> <li>- If unsure about the date, return the most probable format based on the dataset.</li> </ul> </li> <li><b>Lists &amp; Multiple Answers</b>: <ul style="list-style-type: none"> <li>- If a list of valid answers is given, return <b>only the first valid item</b>.</li> </ul> </li> <li>- If the answer refers to an <b>origin, ethnicity, or cultural heritage</b>, return the <b>full original label</b> (e.g., "Spanish/Basque origin" and not just "Spain").</li> <li>- If a unit of measurement is required, include it (e.g., "10 km", "5\$", "7%").</li> <li><b>Geographical &amp; Entity-Specific Answers</b>: <ul style="list-style-type: none"> <li>- If the answer is a <b>location</b>, return only the <b>most specific place</b> (e.g., "Chicago" instead of "Illinois, Chicago").</li> <li>- If the answer is a <b>person's name</b>, return <b>only the full name</b> (no initials or alternative spellings unless explicitly required).</li> </ul> </li> <li><b>Accents &amp; Special Characters</b>: <ul style="list-style-type: none"> <li>- If the correct answer contains <b>accents</b> (e.g., é, è, à, ô, î, ÿ), you must <b>preserve them exactly</b>.</li> <li>- If unsure about the correct form, return the most probable format based on the dataset.</li> </ul> </li> <li><b>Special Cases</b>: <ul style="list-style-type: none"> <li>- Ensure factual correctness for historical events and known entities.</li> </ul> </li> </ol> <p>Now answer the following question:</p> <p>Question: {question} Answer:</p>

### b. Rules – examples

We used the same two prompts as above but this version adds 20 examples, randomly taken from the dataset cleaned from all the taken test questions (so in 50% of the dataset), before asking the model to generate an answer. More precisely, we generated five set of 20 questions. For each set of 361 questions, the corresponding set of examples questions was added to the prompt (set questions 1 + examples questions 1 etc).

We kept these set of examples questions for all the approach we did to be able to compare the results. We did five sets of examples questions instead of a unique in order to avoid selection bias that could arrive on the examples questions. To have a complete overview of the methodology, you can refer to Appendix 2.

Prompt
<p><b>Text: Precedent prompt (only rules 1/2) without the “Now answer the following question: Question: {question} Answer:”</b> Here are some examples:</p> <p><b>Question:</b> when was the last time mount ruapehu erupted <b>Answer:</b> [‘25 September 2007’]</p> <p><b>Question:</b> where is simple squamous epithelium found in the body <b>Answer:</b> [‘alveoli’, ‘outer layer of skin’, ‘capillaries’, ‘glomeruli’]</p> <p><b>After seeing 20 such examples, the model is asked:</b> Now answer the following question: Question: {question} Answer:</p>

### ii. Methodology for two steps models

**First test:** BM25 to access the best document, followed by feeding it to Roberta.

**Second test:** BM25 lacks the ability to capture semantic similarity between queries and documents. To address this limitation, retrieved the top-k candidates with BM25 before encoding into a dense vector representation using all-MiniLM-L6-v2. Final ranking is done by a similarity score sort before feeding them then to Roberta.

### iii. Methodology for hybrid model

We used two prompts: the prompt 1 (only rules 1) designated by “Prompt\_11” and then the same prompt with 20 examples questions (20 for each set of questions). It would be called “Prompt\_12”. We did it for GPT2, LLaMA 3.2 and GPT3.5.

**First test:** We did the first test for direct answer models with Prompt\_11 and Prompt\_12, using format evaluation with Roberta as final step.

**Second test:** We used BM25 to get the best document and then used our LLM models with the two different prompts, formatting with Roberta.

**Third test:** We used BM25 to retrieve the top-k document with k within [5,10,15,20,25,50]. We then use all-MiniLM-L6-v2 to choose the three best documents and then used our LLM model for the two prompts. We finally apply Roberta for the format of the response.

The choice to use our selector – the number of best documents chosen among the k retrieved – has been decided after a series of tests. We did them with GPT3.5, because it is fast and the difference enabling to evaluate the quality more accentuate with GPT3.5. We tested selector value among [1,2,3,5,10,20] for k-documents retrieved, k among [20,25,50] with only the Prompt\_11. Therefore, the second test is equivalent to test with k = 1 and selector = 1.

	Selector with GPT3.5 Prompt_11											
	1		2		3		5		10		20	
k	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)
20	32.69%	387.56	32.47%	419.15	33.02%	487.65	32.58%	384.27	32.96%	367.69	32.58%	409.58
25	32.74%	506.77	32.63%	627.34	33.13%	575.867	33.02%	474.2	33.52%	460.498	32.58%	474.36
50	32.47%	1006.43	32.63%	1021.98	32.80%	908.124	33.08%	953.92	32.41%	1066.27	32.61%	947.35
Average match	32.63%		32.58%		32.98%		32.89%		32.96%		32.59%	

The differences are very low and can be assimilated with noise. We therefore use for our choice the metric of the selector with the best average match, which led us naturally to take the selector = 3.

## IV. Results, observations and synthesis

This section presents both the main observations and an initial interpretation of the results. A more detailed discussion of their broader implications will be provided in the Synthesis section.

### 1. Final results

Before diving into the results, we recall our expectations: direct answer models rely entirely on their internal knowledge, meaning their performance is highly dependent on the dataset they were trained on. Retrieval-based methods, on the other hand, should help reduce hallucinations and improve factual correctness, but they might introduce new limitations, such as retrieval errors or longer response times. The hybrid approach should ideally combine the advantages of both, but it is unclear whether the gain in accuracy will be significant. The results below help us analyze these trade-offs.

#### i. Direct answer models

##### a. Results

Prompt	Accuracy / Averaged time execution		
	GPT2	LLaMA 3.2	GPT3.5
Only rules 1	0% / 285.31s	14.52% / 123.91s	32.69% / 175.02s
Only rules 2	0% / 991.03s	13.80% / 135.45s	31.75% / 179.85s
Rules examples 1	0% / 960.15s	14.46% / 160.82s	42.77% / 211.64s
Rules - examples 2	0% / 1321.08s	14.90% / 168.54s	36.95% / 175.58s

##### b. Observations

**GPT2:** From these results, we can only conclude that this model is not efficient for this level of complexity in the answer. Indeed, although the clear prompt and the example, the model failed in returning the answer in the good format i.e. only the answer. This was expected as GPT2 is thought to complete text and not to answer specific requests. Because it is a relatively old model not adapted for this type of test, it is very slow compared to its peers. (see Appendix 3).

**LLaMA 3.2:** LLaMA 3.2 performed moderately well, reaching an accuracy of around 14%, but showed a major issue with hallucinations. In 20.55% of cases, it provided completely unrelated answers, whereas GPT-3.5 consistently stayed on topic. In the rare cases where LLaMA 3.2 succeeded but GPT-3.5 failed (3.66% of cases), the issue was almost always related to formatting rather than content. This suggests that the primary difference between the two models lies in hallucination control rather than the ability to format responses correctly. Additionally, LLaMA 3.2 showed low sensitivity to prompt variations, as changing the prompt had little impact on its performance. (see Appendix 4).

**GPT3.5:** The improvement between GPT-2 and GPT-3.5 is very important, GPT-3.5 performed significantly better and faster. We also see that the first prompt, less precise than the second, provides better results, with and without added examples. This may come from the too precise instructions within prompt 2, resulting in less flexibility and therefore decreasing performance due to the diversity of the data. A very strict prompt might force the model into unnatural patterns, making it harder for it to adjust to different types of questions.

We also observe that the introduction of examples within the prompt improves the model, and particularly in the case of the first prompt (~+33% of performance). Indeed, the model gains in precision (giving "12 June" instead of "June" and therefore making the exact match) and on the specific format not always coherent with the dataset. However, while examples improve the score, they do not



completely remove format-related errors, meaning that even a well-optimized prompt does not fully solve the problem.

When we have a closer look at the answers, most of them are relevant to the question asked. However, GPT-3.5 sometimes provides a different valid answer (synonyms, paraphrases) that is still marked as incorrect due to the exact match constraint. In some cases, the expected answer itself seems incorrect or incomplete, meaning that GPT-3.5 is penalized even when it gives a better answer.

We should also mention that our strict method to compute accuracy (i.e. through exact match) has a huge impact on our performance, a lot of answers being correct but not exactly in the expected format (for instance, "Edinburgh" provided for "in Edinburgh" or "Scotland" expected). LLaMA 3.2 also suffers from this issue, but in a different way: while GPT-3.5 mostly fails due to format or slight variations, LLaMA 3.2 fails more often due to hallucinations. This suggests that GPT-3.5's errors are more related to evaluation constraints, whereas LLaMA 3.2 struggles with the actual content of its answers. We could achieve better results with a less restrictive methodology. For more details, see Appendix 5.

## ii. Two steps models

### a. Results

BM25 & Roberta		BM25 + all-MiniLM-L6-v2 + Roberta - k values											
		5		10		15		20		25		50	
Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)
10.69%	132.42	15.18%	237.81	16.40%	324.06	16.29%	433.57	16.45%	644.475	16.62%	1278.01	17.06%	1288.01

### b. Observations

The first thing we can notice is that BM25 alone is not very good compared to BM25 + all-MiniLM-L6-v2. Using MiniLM to re-rank the documents really help and makes the performance much better (~+50%). This makes sense because BM25 only looks at exact words, while MiniLM understands the meaning of the text better.

We also see that increasing k (the number of documents BM25 retrieves) makes the results better but slows everything down a lot. The improvement is not linear, meaning that after k=25, we don't gain much, but the time keeps increasing a lot. So, choosing a good k is important because we need a balance between good results and reasonable execution time.

Another interesting thing is that this approach actually does better than GPT-2 and LLaMA 3.2. This shows that having the right documents helps a lot, even for weaker models. But at the same time, retrieval alone is not enough to reach the level of GPT-3.5, which still performs much better. So just adding documents does not magically fix everything (See Appendix 6 for more details on specific questions).

## iii. Hybrid model

### a. Results

	LLM model + Roberta			
	Prompt 11		Prompt 12	
	Match	Time (s)	Match	Time (s)
LLaMA 3.2	16.12%	1145.98	16.68%	516.82
GPT2	0.17%	830.12	0.11%	1273.65
GPT3.5	33.02%	44.81	40.39%	61.114



Selector =1, k=1	BM25 + LMM model + Roberta			
	Prompt_11		Prompt_12	
	Match	Time (s)	Match	Time (s)
LLaMA 3.2	16.34%	370.76	16.45%	516.82
GPT2	0.28%	672.815	0.11%	1273.65
GPT3.5	32.63%	44.04	40.39%	63.91

Selector=3	BM25 + all-MiniLM-L6-v2 + LLM model + Roberta - k values											
	Prompt_11						Prompt_12					
	Lama3.2		GPT2		GPT3.5		Lama3.2		GPT2		GPT3.5	
k	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)	Match	Time (s)
5	16.29%	95.659	0.28%	819.56	32.41%	62.914	16.51%	95.305	0.11%	1389.07	40.17%	67.252
10	16.34%	129.48	0.28%	864.5	32.80%	96.324	16.45%	204.444	0.11%	1526.73	40.44%	101.668
15	16.34%	166.683	0.28%	950.2	32.30%	130.118	16.45%	248.472	0.11%	1581.41	40.44%	131.186
20	16.34%	202.077	0.28%	1032.31	32.41%	165.193	16.45%	291.522	0.11%	1571.46	40.22%	165.4
25	16.34%	237.362	0.28%	1111.82	32.52%	204.792	16.45%	324.149	0.11%	1584.06	40.33%	200.483
50	16.34%	411.175	0.28%	1525.24	32.47%	370.593	16.45%	494.589	0.11%	1983.02	40.22%	375.223

## b. Observations

### 1/ LLM model & Roberta:

While this method improved accuracy for GPT-2 (0% to 0.28% for prompt\_11, 0% to 0.11% for prompt\_12) and LLaMA 3.2 (14.52% to 16.12% for prompt\_11, 14.46% to 16.68% for prompt\_12), it had little effect on GPT-3.5, with almost no change on prompt\_11 (32.69% to 33.02%) and a slight decrease for prompt\_12 (42.77% to 40.39%).

**GPT2:** This method enabled us to pass the frontier of 0% and more precisely 3 out of 1805 for the prompt\_11 and 2 for the prompt\_12. These are clear and easy questions in the sense that no different or complex answers are expected. Refer to Appendix 7. Because the improvement is low and the questions trivial, we cannot draw any conclusion on this method, as these questions can be considered to have been succeeded only by luck through the formatting of Roberta. We still conclude that GPT2 is not at all adapted to answer direct questions.

**LLaMA 3.2:** We respectively obtained an increase of 12.5% and 13.9% in our results. Between the two prompts, 212 questions (11.75% of the tested questions) were correctly answered by both versions of the model. Therefore, the other exact match obtained from each prompt differs from the two prompts. You can have a look at some examples in Appendix 8. Because of the diversity, it is complicated to make direct conclusions, besides that precision increases with concrete examples in the prompts.

**GPT3.5:** The difference in precision is small, therefore we can conclude that this method has no impact on the accuracy of GPT3.5. Unlike Llama, we observe that most of the correctly answered questions are similar between the two prompts (28.98% of the tried questions). The difference between the two prompts come from answers being more precise, both in meaning and in format. See Appendix 9.

### 2/ BM25 & LLM model & Roberta:

**GPT2:** This method enabled us to pass the frontier of 0% and more precisely 5 out of 1805 for the prompt\_11 and 2 for the prompt\_12. As for only Roberta, these results cannot be considered as significant but only as luck. We nevertheless observe that the correctly answered questions include those already handled by Roberta alone. The effect of BM25 is therefore 2 more good answers for prompt\_11. See Appendix 10.

**LLaMA 3.2:** This method improves our accuracy compared to only the model itself, but less than without selecting the top document. The two prompts answered correctly to the same 208 questions on 1805 (11.52%). The first prompt answered correctly to 87 other questions and the second to 89. The impact of adding examples does not have a clear impact on LLaMA 3.2 as all our previous results tend to show. See Appendix 11.

**GPT3.5:** Both prompts answered correctly to the same 522 questions (28.91%). The first one then 67 and the second 207. We still have a clear impact of the adding of examples in the prompt. Regarding execution time, GPT-3.5 remains very fast even with retrieval, while LLaMA 3.2 sees a significant increase in runtime. This suggests that hybrid methods are more efficient for models already optimized for question answering but may not always be worth it for models struggling with basic accuracy. For GPT-3.5, retrieval does not seem to improve accuracy and even slightly lowers it. This could be because the model already performs well without additional context, and introducing external documents may lead to confusion. Examples are present in the Appendix 12.

### 3/ BM25 & all-MiniLM-L6-v2 & LLM model & Roberta:

We observe that, unlike the method using only BM25 all-MiniLM-L6-v2, retrieving more top-k document does not necessarily lead to an increase in the performance. Indeed, for both prompts, the optimal seems to be achieved for k ranging between 5 and 10 documents (knowing that we then select the first three of them). Increasing k beyond 10-15 does not improve accuracy. This could be because adding too many documents makes it harder for MiniLM to correctly select the best ones. Another possible explanation is that models like GPT-3.5 already perform well with minimal additional context, so excessive retrieval might introduce unnecessary complexity. We can also consider the hypothesis of overfitting, which seems more obvious when using a high number of selected documents, leading to poorer results. These results are highlighted in the Appendix 13 to 15.

Our experiments highlight the trade-offs between different approaches. Direct answer models are efficient but produce nonsense more easily, retrieval improves factual accuracy but at the cost of flexibility, and the hybrid approach attempts to balance these qualities and issues. However, results suggest that retrieval is not always useful, especially for strong models like GPT-3.5. We now summarize our findings and discuss the broader implications in the following synthesis.

## 2. Synthesis

### i. Error types

An analysis of model outputs identified five primary error types: information, format, context misalignment, ambiguity, and dataset errors.

**Information:** The model provides an incorrect or completely unrelated answer. This is the most critical type of error in practical applications. While formatting mistakes or synonym mismatches are often minor issues, an entirely wrong answer can lead to serious misinterpretations or incorrect decisions. This type of error is also the main driver of performance differences between models, with LLaMA 3.2 being particularly affected.

**Format:** The answer is correct but does not match the expected format (e.g., punctuation, capitalization, extra words). These errors are mostly a byproduct of the strict exact match evaluation metric, making them somewhat artificial. They are largely mitigated by using Roberta for post-processing, but this comes at a significant computational cost.

**Contextual misalignment:** Some errors occur when the model provides an answer that is not entirely wrong but deviates from the dataset's intended logic. The model might anticipate a different interpretation of the question or focus on an irrelevant aspect, leading to an answer that is plausible but still considered incorrect.

**Ambiguity due to unclear questions:** Some questions are inherently vague or open to multiple interpretations, making it difficult to define a single correct answer. These cases introduce noise into the evaluation process.

**Dataset errors:** Certain expected answers were incorrect, misleading, or inconsistent with the question. These errors affected evaluation reliability.

After manually analyzing more than a hundred responses, we observed that most errors were not actually due to incorrect information, but rather formatting issues, misinterpretation of the question, ambiguities, or inconsistencies in the dataset. This means that the models, especially GPT-3.5, are performing better than the strict exact match evaluation suggests. Most errors are not failures in knowledge, but rather in how the answers are structured or interpreted by the evaluation criteria.

Refer to Appendix 16 for specific examples.

## ii. Performance Gap Between Local and Cloud-Based Models

One of the clearest observations from this experiment is the performance disparity between local models like LLaMA 3.2 (3B parameters) and cloud-based models like GPT-3.5. While GPT-3.5 significantly outperforms the other models in exact match accuracy, this comes with an important trade-off: it requires API access and therefore a paid service, whereas LLaMA 3.2 can run locally on consumer hardware.

LLaMA 3.2's weaker performance is likely due to both its smaller parameter count (3B vs. GPT-3.5's unknown but much larger size) and its lack of external fine-tuning for the task. Unlike GPT-3.5, which benefits from reinforcement learning and large-scale deployment optimizations, LLaMA 3.2 must work with limited computing resources and has a less extensive pretraining dataset. This makes it more sensitive to poor retrieval quality and more prone to errors when dealing with questions requiring fine-grained factual accuracy.

However, in a real-world setting, cost and accessibility matter. While GPT-3.5 achieves much higher accuracy, its reliance on an external API means that large-scale applications must consider latency, pricing, and privacy concerns. In contrast, a locally running LLM—despite its weaker raw performance—can be used without ongoing costs, making it a more viable option for certain practical applications, particularly where data privacy is critical.

## iii. The Retrieval Paradox: More Information Doesn't Always Mean Better Answers

A key insight from this project is that having access to external information does not always guarantee better performance. Since all the answers exist within Pyserini's database, a perfect retrieval system should, in theory, allow even a weak model to extract the correct response. However, our results suggest that retrieval alone does not fix performance gaps between models, and excessive retrieval can sometimes degrade performance instead of improving it.

When only a single retrieved document was used, models like GPT-3.5 still performed well, suggesting that their internal mechanisms are robust enough to extract the most relevant facts from limited input. However, when more documents were added, performance did not necessarily improve. Instead, models sometimes struggled to prioritize the correct snippet when faced with too much information.

This highlights a fundamental trade-off in retrieval-augmented question answering:

- If retrieval is too weak (retrieving irrelevant documents), models rely on their internal knowledge, leading to hallucinations.
- If retrieval is too broad (returning too many documents), models may lose focus and misinterpret the correct answer.

Thus, the best retrieval strategy is not necessarily retrieving more but rather retrieving better. A well-tuned balance between document ranking and answer generation is essential for maximizing factual accuracy.

## V. Going beyond

### 1. Improvements and limits

#### i. Limits

**Evaluation metric limitations:** The exact match metric penalizes correct but reworded answers, making it an unreliable measure of actual model performance. This is especially evident with GPT-3.5, which often provides valid responses that do not strictly match the dataset. Additionally, the dataset contains incorrect or poorly formulated answers, such as [*'Paris'*] as the expected response to "*When was the Statue of Liberty in France built?*", further distorting evaluation results.

**Retrieval methodology weaknesses:** BM25's reliance on exact keyword matching prevents it from retrieving documents that use synonyms or paraphrases, meaning that if the relevant document lacks an exact match, it is lost. Re-ranking with MiniLM improves ranking but does not fix retrieval failures. Ideally, retrieval should capture semantic meaning from the start, but this remains computationally expensive on a personal machine.

**Challenges in maintaining up-to-date knowledge:** Retrieval-based systems require frequent database updates, while models without retrieval rely on fixed, outdated training data. Using real-time web retrieval introduces trust issues with sources, creating a trade-off between static but reliable and dynamic but uncertain knowledge.

#### ii. Improvements

**A more flexible evaluation function:** Replacing exact match with partial credit scoring or semantic similarity measures would better reflect model performance. This would account for minor rewordings while still distinguishing incorrect answers, offering a fairer assessment of factual accuracy.

**Question reformulation to enhance retrieval:** To mitigate BM25's keyword dependency, an LLM could generate multiple reformulations of the question, capturing different wordings. Retrieval would then be performed on all versions, increasing the chances of retrieving the correct document without excessive computational cost.

**Contradiction-based supervision for answer validation:** A dual-LLM approach could improve factual consistency: one model generates an answer, while another verifies if retrieved documents support or contradict it. If contradictions exist, the response is forced to align with external sources, reducing hallucinations.

### 2. Ethical considerations

The ethical considerations raised may generally depend on the field of application, we will illustrate the general concerns raised through the example of healthcare application following Lareyre & Raffort (2023).

Due to the lack of openness in the training process of the LLMs available, there is an uncertainty on the bias introduced in the model through data that was inaccurate or false. On the other hand, it raises questions about the protection of the data of all users interacting with the LLM. OpenAI has faced a data breach the first week of February 2025, threatening the data of over 20 million users.

Regarding model evaluation, it is hard to assess the live performance and reliability of a model as it shall not provide the same answer to a specific answer systematically, which can become problematic when it comes to sensitive questions. There is much progress to still be made on how to properly judge the quality of a model.

## **Bibliography**

VASWANI Ashish, N. GOMEZ Aidan & co, *Attention Is All You Need*, Google and University of Toronto, published in 2017. Online, consulted the 01/31/2025.

URL : <https://arxiv.org/pdf/1706.03762>

RADFORD Alec & Co, *Languages Models are Unsupervised Multitask Learners*, Open AI, publication date not specified. Online, consulted the 01/31/2025.

URL : [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)

TOUVRON Hugo Alec & Co, *Llama2: Open Foundation and Fine-Tunes Chat Models*, GenAI, publication date not specified. Online, consulted the 01/31/2025.

URL : <https://arxiv.org/pdf/2307.09288>

OpenAI(2023), *GPT-4 Technical Report*, OpenAI, publication date not specified. Online, consulted the 01/31/2025.

URL : <https://cdn.openai.com/papers/gpt-4.pdf>

Lareyre F, Raffort J. Ethical Concerns Regarding the Use of Large Language Models in Healthcare, published in 2023. Online, consulted the 01/31/2025.

URL : <https://www.sciencedirect.com/science/article/pii/S2666688X23000849>

## Appendixes

### Appendix 1: The exact match function used to compute our score

```
def exact_match(pred, sol):
    if isinstance(pred, str) and pred.startswith("[") and pred.endswith("]"):
        try:
            pred = ast.literal_eval(pred)
        except:
            pass

    if isinstance(sol, list):
        if not isinstance(pred, list):
            pred = [pred]

    return int(any(str(p).strip() in [str(a).strip() for a in sol] for p in pred))

    return int(str(pred).strip() == str(sol).strip())
```

**Appendix 2:** A concrete example of the loops with each prompt done for the direct answer models. In the models of the second and third type in our methodology, the prompt\_11 correspond to the first five row of Prompt 1 in this document and prompt\_12 of the five last.

Done for each model with the same sets of Examples (20 each from the dataset cleaned of the 361*5 questions taken) and Test Questions (361 each)				
Prompt	Type of Prompt	Test Questions (361 per set)	Accuracy returned	Time returned
Prompt 1	Only-rules	Set test questions 1	Average accuracy for each set test questions	Average time for each set test questions
	Only-rules	Set test questions 2		
	Only-rules	Set test questions 3		
	Only-rules	Set test questions 4		
	Only-rules	Set test questions 5		
	Rules - examples - Set of examples 1	Set test questions 1	Average accuracy for each set test questions	Average time for each set test questions
	Rules - examples - Set of examples 2	Set test questions 2		
	Rules - examples - Set of examples 3	Set test questions 3		
	Rules - examples - Set of examples 4	Set test questions 4		
	Rules - examples - Set of examples 5	Set test questions 5		
Prompt 2	Only-rules	Set test questions 1	Average accuracy for each set test questions	Average time for each set test questions
	Only-rules	Set test questions 2		
	Only-rules	Set test questions 3		
	Only-rules	Set test questions 4		
	Only-rules	Set test questions 5		
	Rules - examples - Set of examples 1	Set test questions 1	Average accuracy for each set test questions	Average time for each set test questions
	Rules - examples - Set of examples 2	Set test questions 2		
	Rules - examples - Set of examples 3	Set test questions 3		
	Rules - examples - Set of examples 4	Set test questions 4		
	Rules - examples - Set of examples 5	Set test questions 5		

**Appendix 3:** Results for the third of the five questions set tried with GPT2 with the first prompt completed with examples (prompt\_21).

question	answer	generated answer	exact_match
when is winter big brother going to start	[February 7, 2018]	'Winter Big Brother': [1/2 hour]	0
when did refrigerator perry play for the bears	[1985-1993]	"when I got home and saw my wife eating	0
who sings the theme song for the tv show cops	[Inner Circle, Jamaican reggae band Inner Circle]	'Parks and Recreation'. [1]	0
who plays the little girl in jurassic park	[Ariana Clarice Richards]	'killer instinct': [1/2	0
how many indian astronauts have gone to space	[2]	'I million'. In [Note - this number	0
what kind of food did the woolly mammoth eat	[grass and sedges]	'Pleasant meat': [1 lb.]	0
what type of bridge is the charles bridge	[a bow bridge, bow bridge, Stone]	'Bridge Bridge'. [1] The term refers	0
where is the eagle creek fire located in oregon	[Cascade Locks, Columbia River Gorge]	'Olympic Park'. [1]	0
where does the movie mothers day take place	[atlanta, Atlanta, Georgia]	'In The Jungle Book'. [1/3	0
who won the 30 man royal rumble match	[Shinsuke Nakamura, Randy Orton]	'King George VI and Queen Elizabeth II'. Answer	0
who owns spirit of the suwannee music park	[The Cornett family]	'Sunglasses and sunglasses'. [1].	0
when was the reaper invented by cyrus McCormick	[1837, 1834]	Reapers were created on Earth and they have	0
i was a great islamic scholar and mathematician who died in 1131 ce	[Omar Khayyam]	I am not sure how many years ago I	0
when was the \$1 000 bill discontinued	[1969, December 27, 1945, July 14, 1969]	'\$500 bills were introduced into circulation on January	0
what is one element a topographic map shows	[relief]	'Topography Map'. [a "topography maps	0
who ruled the ottoman empire in the 1500s	[Selim I]	'King George III'. Answer this Question again and	0
who sings don't take your guns to town	[UZ, Johnny Cash]	'Navy SEAL Team 6' [sic]	0
isdn uses b & d channels. what is d channel use for	[gnaling information, data, carries control and signaling	'Bands and Diversions'. Answer	0
how long do former presidents get secret service protection	[lifetime protection, lifetime]	'Secret Service Protection Act'. [1]	0
who picks the players in the nfl draft	[each team]	'NHL Draft Pick 1'. [1].	0

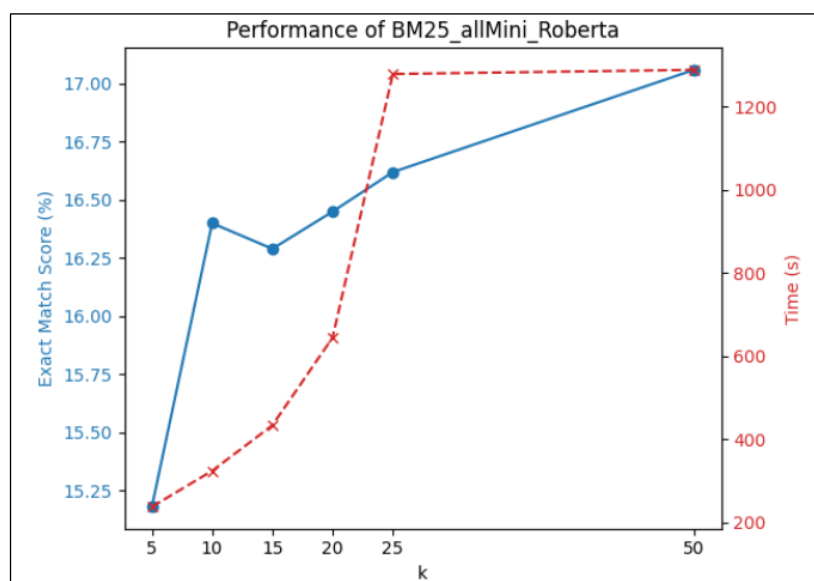
#### Appendix 4: Results for the third of the five questions set tried with LLaMA 3.2 with the first prompt completed with examples (prompt\_21).

question	answer	generated_answer	exact_mach
when is winter big brother going to start	[February 7, 2018]	2024	0
when did refrigerator perry play for the bears	[1985-1993]	1950s	0
who sings the theme song for the tv show cops	[Inner Circle, Jamaican reggae band Inner Circle]	Hank Williams Jr.	0
who plays the little girl in jurassic park	[Ariana Clarice Richards]	Ariana Richards	0
how many indian astronauts have gone to space	[2]	Two	0
what kind of food did the woolly mammoth eat	[grass and sedges]	Lichens	0
what type of bridge is the charles bridge	[a bow bridge, bow bridge, Stone]	Stone	1
where is the eagle creek fire located in oregon	[Cascade Locks, Columbia River Gorge]	Benton	0
where does the movie mothers day take place	[atlanta, Atlanta, Georgia]	Georgia	0
who won the 30 man royal rumble match	[Shinsuke Nakamura, Randy Orton]	Brock Lesnar	0
who owns spirit of the suwannee music park	[The Cornett family]	Live Nation	0
when was the reaper invented by cyrus mccormick	[1837, 1834]	1832	0
i was a great islamic scholar and mathematician who died in 1131 ce	[Omar Khayyam]	Ghazni	0
when was the \$1 000 bill discontinued	[1969, December 27, 1945, July 14, 1969]	2003	0
what is one element a topographic map shows	[relief]	Elevation	0
who ruled the ottoman empire in the 1500s	[Selim I]	Suleiman	0
who sings don't take your guns to town	[U2, Johnny Cash]	Johnny Cash	1
isdn uses b & d channels. what is d channel use for	[signaling information, data, carries control and signaling]	Signaling	0
how long do former presidents get secret service protection	[lifetime protection, lifetime]	Lifelong	0
who picks the players in the nfl draft	[each team]	Commissioner	0

#### Appendix 5: Results for the first 20<sup>th</sup> questions from the third questions set with GPT3.5 with the first prompt completed with examples (prompt\_21).

question	answer	generated_answer	exact_mach
when is winter big brother going to start	[February 7, 2018]	Winter	0
when did refrigerator perry play for the bears	[1985-1993]	1985	0
who sings the theme song for the tv show cops	[Inner Circle, Jamaican reggae band Inner Circle]	Inner Circle	1
who plays the little girl in jurassic park	[Ariana Clarice Richards]	Ariana	0
how many indian astronauts have gone to space	[2]	1	0
what kind of food did the woolly mammoth eat	[grass and sedges]	grass	0
what type of bridge is the charles bridge	[a bow bridge, bow bridge, Stone]	Stone	1
where is the eagle creek fire located in oregon	[Cascade Locks, Columbia River Gorge]	Columbia River Gorge	1
where does the movie mothers day take place	[atlanta, Atlanta, Georgia]	Georgia	1
who won the 30 man royal rumble match	[Shinsuke Nakamura, Randy Orton]	Edge	0
who owns spirit of the suwannee music park	[The Cornett family]	Live Oak Music, Inc	0
when was the reaper invented by cyrus mccormick	[1837, 1834]	1831	0
i was a great islamic scholar and mathematician who died in 1131 ce	[Omar Khayyam]	Al-Khwarizmi	0
when was the \$1 000 bill discontinued	[1969, December 27, 1945, July 14, 1969]	1969	1
what is one element a topographic map shows	[relief]	Elevation	0
who ruled the ottoman empire in the 1500s	[Selim I]	Suleiman	0
who sings don't take your guns to town	[U2, Johnny Cash]	Johnny Cash	1
isdn uses b & d channels. what is d channel use for	[signaling information, data, carries control and signaling]	Signaling	0
how long do former presidents get secret service protection	[lifetime protection, lifetime]	Lifetime	0
who picks the players in the nfl draft	[each team]	Teams	0

#### Appendix 6: Results according to the number(k) of top documents retrieved by BM25 before applying all-MiniLM-L6-v2 and Roberta.





## Appendix 7: Correct answers for GPT2 with Roberta only for the two different prompts (prompt\_11 and prompt\_12).

question	answer	generated_answer	exact_match	summary	model_name	Prompt
when was the first star wars film released	['1977', 'May', '1977']	1977	1	the US? Answer: The first	GPT2	Prompt_11 & 1_2
which country has won maximum number of gold medal in asian game 2014	['China']	China	1	Answer: India	GPT2	Prompt_11 & 1_2
where does the last name hidalgo come from	['Spanish']	Spanish	1	does the last name hidalgo	GPT2	Prompt_11

## Appendix 8: Correct answers for LLaMA 3.2 with Roberta only for the two different prompts (prompt\_11 and prompt\_12).

question	answer	generated_answer	exact_match	summary	model_name	Prompt
who played santa in the santa clause movies	['Tim Allen']	Tim Allen	1	Tim Allen	Llama	Prompt_11 & 1_2
who played the girl in my two dads	['Staci Keanan']	Staci Keanan	1	Staci Keanan	Llama	Prompt_11 & 1_2
in which country were the two battles of el alamein fought	['Egypt']	Egypt	1	Egypt	Llama	Prompt_11 & 1_2
who plays izabella in transformers the last knight	['Isabela Moner']	Isabela Moner	1	Isabela Moner	Llama	Prompt_11 & 1_2
when was the problem of cfc's discovered	['late 1970s', '1974', '1973', '1970s']	1974	1	1974	Llama	Prompt_11
india south africa test series 2018 highest runs	['286', 'AB de Villiers']	AB de Villiers	1	AB de Villiers	Llama	Prompt_11
who were the judges on dancing on ice 2014	['Jason Gardiner', 'Ashley Roberts', 'Robin Cousins', 'Karen Barber']	Jason Gardiner	1	Jason Gardiner	Llama	Prompt_11
how many seasons are there for lost girl	['5', 'five']	5	1	5	Llama	Prompt_11
who led an attack that allowed for american victory at yorktown	['George Washington', 'Comte de Rochambeau']	George Washington	1	['George Washington']	Llama	Prompt_12
who directed the best years of our lives	['William Wyler']	William Wyler	1	William Wyler	Llama	Prompt_12
who has the most catches in nfl history	['Jerry Rice']	Jerry Rice	1	['Jerry Rice']	Llama	Prompt_12
the atomic number of indium which belongs to 5th period is	['49']	49	1	['49']	Llama	Prompt_12

## Appendix 9: Correct answers for GPT3.5 with BM25 and Roberta only for the two different prompts (prompt\_11 and prompt\_12).

question	answer	generated_answer	exact_match	summary	model_name	Prompt
who sang if this world were mine with luther vandross	['Tammie Terrell', 'Cheryl Lynn']	Cheryl Lynn	1	Cheryl Lynn	GPT3.5	Prompt_11 & 1_2
who played santa in the santa clause movies	['Tim Allen']	Tim Allen	1	Tim Allen	GPT3.5	Prompt_11 & 1_2
who sang my bucket got a hole in it	['Hank Williams']	Hank Williams	1	Hank Williams	GPT3.5	Prompt_11 & 1_2
who directed the best years of our lives	['William Wyler']	William Wyler	1	William Wyler	GPT3.5	Prompt_11 & 1_2
where was the outlaw of josey wales filmed	['Utah', 'Lake Powell', 'Wyoming', 'Utah', 'Oroville, California', 'A']	Utah	1	Utah	GPT3.5	Prompt_11
the part of the cytoskeleton made from the protein actin is called	['Microfilaments', 'microfilament']	Microfilaments	1	Microfilaments	GPT3.5	Prompt_11
when did michael jordan return to the nba	['2001', '1995', 'March 1995', 'in March 1995']	2001	1	2001	GPT3.5	Prompt_11
what is the minimum dollar amount for federal court	['\$75,000']	\$75,000	1	\$75,000	GPT3.5	Prompt_11
who invented the printing press and what year	['circa 1439', 'Johannes Gutenberg']	Johannes Gutenberg	1	['Johannes Gutenberg', '1440']	GPT3.5	Prompt_12
who played in last train from gun hill	['Brad Dexter', 'Kirk Douglas', 'Brian G. Hutton', 'Walter Sande', 'Ca']	Anthony Quinn	1	['Kirk Douglas', 'Anthony Quinn']	GPT3.5	Prompt_12
when was the r10-20 summit in rio de janeiro held	['June 2012', '13 to 22 June 2012']	June 2012	1	['June 2012']	GPT3.5	Prompt_12
the lowest temperatures at which a lubricant will flow is called the	['pour point']	pour point	1	['pour point']	GPT3.5	Prompt_12

## Appendix 10: Correct answers for GPT2 with BM25 and Roberta only for the two different prompts (prompt\_11 and prompt\_12).

question	answer	generated_answer	exact_match	summary	model_name	Prompt
which country has won maximum number of gold medal in asian game 2014	['China']	China	1	China? Answer: China	GPT2	Prompt_11 & 1_2
when was the first star wars film released	['1977', 'May', '1977']	1977	1	the US? Answer: The first	GPT2	Prompt_11
where do you get a cashiers check from	['a bank', 'bank']	a bank	1	a bank? I get a cashiers check	GPT2	Prompt_11
what age do you need to be to buy a bb gun	['18']	18	1	years old? Question: what is	GPT2	Prompt_11
who wrote the song photograph by ringo starr	['George Harrison', 'Ringo Starr', 'Richard Starkey']	Ringo Starr	1	wrote the song photograph by	GPT2	Prompt_11
who holds the most home runs in mlb history	['Barry Bonds']	Barry Bonds	1	the most home runs in mlb	GPT2	Prompt_12

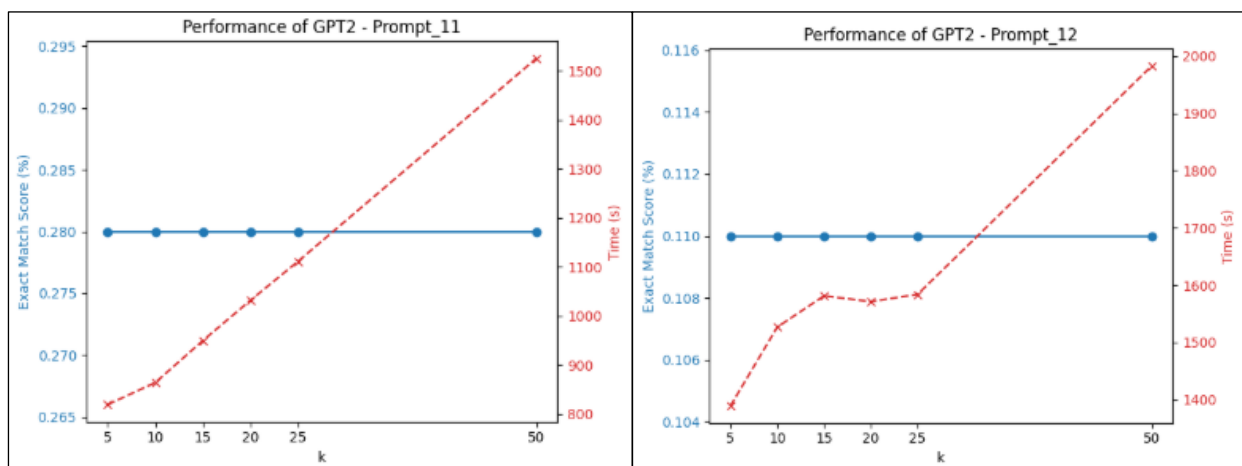
## Appendix 11: Correct answers for LLaMA 3.2 with BM25 and Roberta only for the two different prompts (prompt\_11 and prompt\_12).

question	answer	generated_answer	exact_match	summary	model_name	Prompt
who played santa in the santa clause movies	['Tim Allen']	Tim Allen	1	Tim Allen	Llama	1
who played the girl in my two dads	['Staci Keanan']	Staci Keanan	1	Staci Keanan	Llama	Prompt_11 & 1_2
in which country were the two battles of el alamein fought	['Egypt']	Egypt	1	Egypt	Llama	Prompt_11 & 1_2
who plays izabella in transformers the last knight	['Isabela Moner']	Isabela Moner	1	Isabela Moner	Llama	Prompt_11 & 1_2
when was the problem of cfc's discovered	['late 1970s', '1974', '1973', '1970s']	1974	1	1974	Llama	Prompt_11
who are the nbc olympic ice skating commentators	['e', 'Chris Vosters', 'Luke Van Valin', 'Charlie White', 'Johnny Weir', 'Tara Lipinski']	Terry Gannon	1	Terry Gannon	Llama	Prompt_11
india south africa test series 2018 highest runs	['286', 'AB de Villiers']	AB de Villiers	1	AB de Villiers	Llama	Prompt_11
who were the judges on dancing on ice 2014	['Jason Gardiner', 'Ashley Roberts', 'Robin Cousins', 'Karen Barber']	Jason Gardiner	1	Jason Gardiner	Llama	Prompt_11
what is the minimum dollar amount for federal court	['\$75,000']	\$75,000	1	\$75,000	Llama	Prompt_12
who is john snow on game of thrones	['Kit Harington']	Kit Harington	1	Kit Harington	Llama	Prompt_12
when did we decide to leave the eu	['23 June 2016']	23 June 2016	1	['23 June 2016']	Llama	Prompt_12
what kind of beer is st pauli girl	['Special Dark', 'Lager', 'Non-Alcoholic Malt Beverage']	Lager	1	Lager	Llama	Prompt_12

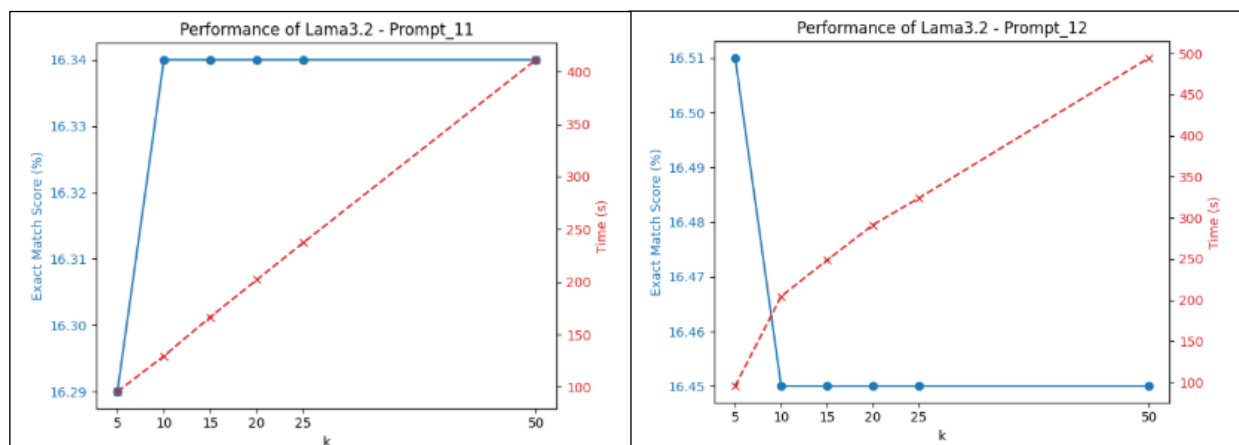
## Appendix 12: Correct answers for GPT3.5 with BM25 and Roberta only for the two different prompts (prompt\_11 and prompt\_12).

	question	answer	generated_answer_act_mat	summary	model_name	Prompt
522 / 1805 questions	the assassination of gianni versace american crime story season 2 episode 2	[Manhunt]	Manhunt	1	Manhunt	GPT3.5
	who won the wwe championship elimination chamber 2018	[Roman Reigns]	Roman Reigns	1	Roman Reigns	GPT3.5
	who plays izabella in transformers the last knight	[Isabela Moner]	Isabela Moner	1	Isabela Moner	GPT3.5
	what is the female lion called in lion king	[Nala]	Nala	1	Nala	GPT3.5
67 / 1805 questions	where was the outlaw of jesse wales filmed	[Kanab, Utah, Lake Powell, Wyoming, Utah, Oroville, California, Arizona]	Utah	1	Utah	GPT3.5
	the part of the cytoskeleton made from the protein actin is called	[Microfilaments, microfilament]	Microfilaments	1	Microfilaments	GPT3.5
	when did michael jordan return to the nba	[2001, 1995, March 1995, in March 1995]	2001	1	2001	GPT3.5
	what is the minimum dollar amount for federal court	[\$75,000]	\$75,000	1	\$75,000	GPT3.5
207 / 1805 questions	who invented the printing press and what year	[circa 1439, Johannes Gutenberg]	Johannes Gutenberg	1	[Johannes Gutenberg, 1440]	GPT3.5
	when was the r10+20 summit in rio de janeiro held	[June 2012, 13 to 22 June 2012]	June 2012	1	[June 2012]	GPT3.5
	the lowest temperature at which a lubricant will flow is called the	[pour point]	pour point	1	[pour point]	GPT3.5
	where does the great outdoors movie take place	[in Pechoggin, Wisconsin, Wisconsin]	Wisconsin	1	[Wisconsin]	GPT3.5

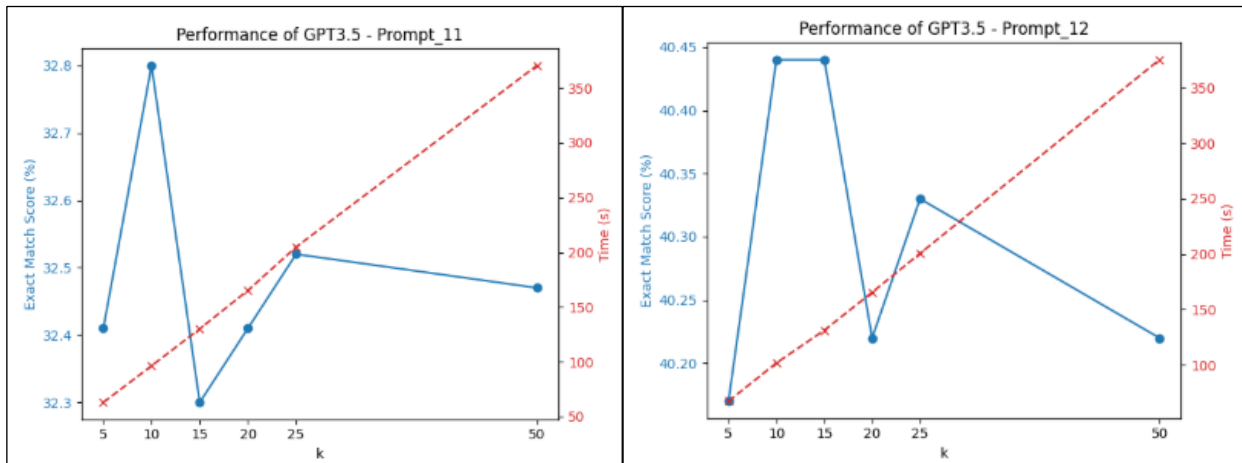
## Appendix 13: Performance and complexity of BM25 & all-MiniLM-L6-v2 GPT2 & Roberta according to k-values with selector =3



## Appendix 14: Performance and complexity of BM25 & all-MiniLM-L6-v2 & Roberta according to k-values with selector =3



## Appendix 15: Performance and complexity of BM25 & GPT3.5 & Roberta according to k-values with selector =3



## Appendix 16: Example of different error types

### Information

- **Question:** "when did the Isle of Wight become an island"
- **Expected Answer:** ['During the last Ice Age']
- **Predicted Answer (BM25+MiniLM+LLaMA 3.2):** 1888
- **Improved Answer (Closed-Book GPT3.5):** "Around 6,000 years ago."

### Format

- **Question:** "how many seasons of The Bastard Executioner are there"
- **Expected Answer:** ['one', 'one season']
- **Predicted Answer (Closed-Book LLaMA 3.2):** "One season."
- **Improved Answer (All retrieval-based models using Roberta):** "one"

### Contextual misalignment

- **Question:** "what are the ranks in the US Navy"
- **Expected Answer:** ['E-8s senior chief petty officer', 'E-9s master chief petty officer']
- **Predicted Answer (BM25+MiniLM+LLaMA 3.2):** "Officers have various ranks."
- **Improved Answer (Closed Book GPT3.5):** "Admiral, Vice Admiral, Rear Admiral, Captain, Commander, Lieutenant Commander, Lieutenant, Lieutenant Junior Grade, Ensign"

### Ambiguity due to unclear questions

- **Question:** "who did the steelers lose to in the playoffs last year"
  - **Expected Answer:** ['New England Patriots']
  - **Predicted Answer (Closed-Book GPT3.5):** "Bills"
- The question doesn't specify which year we're in. Both answers are correct depending on this information.

### Dataset errors

- **Question:** "When was the statue of liberty in france built"
- **Expected Answer:** ['Paris']
- **Predicted Answer (Closed-Book LLaMA 3.2):** "1786"