EquinorQA: Large Language Models for Question Answering over proprietary data

Darío Garigliotti[a], Bjarte Johansen[b], Jakob Vigerust Kallestad[b], Seong-Eun Cho[b] and Cèsar Ferri[c]

[a]University of Bergen, Norway
[b]Equinor, Norway
[c]Universitat Politècnica de València, Spain

Abstract. Large Language Models (LLMs) have become the state-of-the-art technology in a variety of language understanding tasks. Accordingly, many commercial organizations have been increasingly trying to integrate LLMs in multiple areas of their production and analytics. A typical scenario is the need for answering questions over a domain-specific, private collection of documents, such that the answer is supported by evidence clearly referenced from those documents. The Retrieval-Augmented Generation (RAG) framework has been recently used by many applications for this kind of scenarios, as it intuitively bridges dedicated data collections and state-of-the-art generative models. Yet, LLMs are known to present data contamination, a phenomenon in which their performance on evaluation data relevant to a task is influenced by said data being already incorporated to the LLM during training phase. In this paper, we assess the performance of LLMs within the domain of Equinor, the largest energy company in Norway. Specifically, we address question answering with a RAG-based approach over a novel data collection not available for well-established LLMs during training, in order to study the effect of data contamination for this task. Beyond shedding light on LLM performance for a highly-demanded, realistic industrial scenario, we also analyze its potential impact for an ensemble of personas in Equinor with particular information needs and contexts.

1 Introduction

Large language models (LLMs) have become the dominant technology in language processing. These models, which consist of billions of parameters, are trained on vast amounts of general-purpose textual content to address a variety of tasks such as text classification, textual entailment, data wrangling, and question answering (QA) [16, 8, 20]. While most of the prominent LLMs are trained in a broad range of domains, other works focus on building smaller models for specialized domains where they end up outperforming general-purpose LLMs [18]. As LLMs are prone to hallucinating, all across their applications there is an interest for verifying that the claims that occur in a text generated by such a model is truthful [11, 14]. Within the research on QA, this expectation takes shape by several similar problems such as attribution in question answering [4], evidentiality-guided generation [3], verifiability of generation [11], and factuality in summarization [12]. One of them in particular, self-supported question answering (SQA) [14], aims to generate an answer for a question and complement the answer with a document passage that supports that the answer is appropriate. Additionally, there is an argument for supported answers making models more explainable, as they contribute to increase trust in model outputs [14]. In this work, we propose to study SQA within domain-specific LLMs. Specifically, we aim to provide supporting evidence alongside a response for a question, in the domain of energy industry.

The implicit knowledge stored in its numerous parameters often makes a LLM suitable, for example, to answer open-domain questions, although they may become limited without additional explicit knowledge as context where to get the answer from [10]. The retrieval-augmented generation (RAG) paradigm achieves state of the art in QA by augmenting the base LLM with contexts retrieved as relevant for a question and then generating answers accordingly [6].

Our work is an assessment of the ability of LLMs to be extended with non-parametrized knowledge in a domain-specific industrial scenario while avoiding contamination with proprietary data. In this paper, we study strategies to exploit RAG for SQA within closed LLMs. Concretely, we compare possible mechanisms to achieve RAG when using a widely-established LLMs, as base models, within the domain of energy industry at Equinor, the largest energy company in Norway. The basic strategy consists in augmenting the question with the retrieved passages for the question altogether in the input for the LLM at inference time to the base model. Our contributions are as follows: (i) we develop EquinorQA, a novel dataset for studying the phenomenon of contamination in proprietary data; (ii) we analyze the capabilities of LLM-powered, RAG-based methods for SQA under several experimental configurations; (iii) we discuss the implications of assessing and deploying solutions based on these methods for a variety of professional roles in the company with particular information need scenarios.

The remainder of this paper is organized as follows. We first discuss related literature. Then, we define the task of supported QA and describe the construction of the EquinorQA dataset. After that, we detail our approach and research questions. We then proceed to report our experimental results. We conclude with a discussion on the industrial impact of our study and a proposal for future work.

2 Related Work

Novel datasets on tasks closely related to supported question answering confirm the interest and recency of the research community in addressing these scenarios. HAGRID, for example, is a collection of items made of a question and its contexts or passages (annotated by humans as relevant to the question), together with a LLM-generated

answer with citation, and human judgements about whether the answer is correct and attributable from the cited context [9]. Related to HAGRID, MIRACL presents instead a core retrieval dataset, with human annotations about question - document (passage) pairs, for the same documents in several languages [24]. AttrScore is a test collection with items relating a question, answer, passage, and judgment about the passage for the answer (supporting, related, or contradictory) [23]. A recent framework [21] exploits citation-related data to improve new citation-centered generations. A dataset on counterfactual questions recently released, IfQA, [22] only evaluates on GPT methods, while it does not study open LLMs, nor assesses configurations in possible RAG components.

The concern for data contamination in LLMs, this is, the phenomenon where test data has become part of the training phase of a model, has been recently identified [7, 17], responding to the increasing scale of data volume and model size with which they are trained, usually by massive web crawling [5, 2, 15]. A main impact of this phenomenon is the influence of the contaminated data over the model to exhibit a performance better than its actual ability in a particular test set [13, 17]. The detection and measurement of data contamination present crucial challenges for the evaluation of LLMs.

3 Problem Definition and Dataset Construction

We follow a task setup for self-supported QA recently studied in the literature [6]. Given a question and a set of text passages, a method must produce a text output made of statements, answering the question, each statement citing one or more passages from the set.

The exact task to address is determined by further constraints on the allowed questions, passages and outputs, such as criteria regarding (i) what kind of questions and answers are to be involved, (ii) how to split documents into the passages that made the corpus, (iii) how to segment the output into statements, (iv) how many passages are to be cited, and (v) what format the citations are represented in, (vi) how many correct answers there are.

We conduct our experiments on EquinorQA, a novel dataset that was built by experts in the domain with the objective of capturing question-answering phenomena over proprietary data about energy. Although the foundational data used to build it is not private, it is nonetheless of the same nature as the data that is present in the kind of information access problems that we aim to study. Specifically, we use publicly available web articles that report corporate news officially released by Equinor.[1] The test instances that are collected by experts from these articles allow to experiment on supported question answering in the scenario with LLMs free from contamination regarding our dataset. With actual proprietary data, the contamination-free scenario holds since the training stage of a LLM has no access to the data. The kind of information and linguistic style present in these articles appear recurrently in question answering needs in Equinor, not only for factoid replies or more elaborated answers, but also when asking for the stance of the company on a particular topic.

As these articles are released with a frequency of some tens per month, we select a total of 60 articles published in English between August 1st, 2023, and March 8th, 2024. We choose this starting date due to the fact that July 2023 is the latest training phase cut-off among the LLMs that we experiment with.[2] By doing so, we ensure that the selected articles were not part of the training phase for any of the LLMs here studied. After all, our goal is to assess LLMs over data under this constraint, such as Equinor proprietary data.

After obtaining the raw textual content of each article, we post-process it by removing subtitles and splitting it by paragraphs. Each of these paragraphs is a unique passage —or context— in our passage collection. Accordingly, we assign a unique identifier to each passage, with the identifier being a random alphanumerical string of eight characters. Otherwise, as initial experiments showed, if a set of n passages is provided to an LLM in the prompt —as possible contexts where to find an answer and cite a source— and the passage references for citation are $[1], ..., [n]$, this seems to increase the chance that the LLM will hallucinate references in the form of natural numbers that are not among the ones provided.

In next stage, domain experts collect 60 questions and their corresponding answers and relevant passages. Since few articles are rather very short and repetitive —for example, quarterly reporting on stock options—, some articles lead each to multiple questions. The questions formulated by the experts aim to satisfy certain requirements: (i) each question is very similar to the kind of questions actually asked by Equinor employees using QA systems, (ii) each question is rather focused so that its answer is a very short phrase, and (iii) it is focused also to ensure that identifying the relevant passages per question is feasible in a reasonable time. Two domain experts worked on formulating questions within these criteria, and collecting their answers and relevant passages; they then reached total agreement on the instances to be in the final dataset.

Regarding criterion (ii), as shown in Table 1, the correct answer is always one single short phrase. Yet, some questions allow for more than one possible way to express the correct answer, for example, "batteries" and "battery", or "mid-2024" and "mid 2024." We also require to incorporate questions that would allow for further studying the reasoning capabilities of the LLMs. For this, the question set consists of three groups —or types— of 20 questions each:

- *Regular (REG)*: these are questions whose answer is a noun phrase that appears in a provided context.
- *Maybe Not Fresh (MNF)*: although the articles were not seen by a LLM, the answer to a question over them might be present in older articles that are earlier than the training cut-off date. Hence, a MNF question is such its answer might not be fresh.
- *Hard (HARD)*: a hard question often asks about arithmetic operations like counting, summing, or comparing two values; it could instead ask whether an event has had positive outcome for Equinor.

If a question satisfies the characterization of more than one of these groups, it is assigned to the group with highest priority, according to the order HARD $>$ MNF $>$ REG. Table 1 shows examples of questions and answers for each query type. Our dataset and related resources are made publicly accessible in a repository.[3]

4 Approach

In this section we present our question answering approach based on the Retrieval-augmented Generation (RAG) framework [10]. We first describe the overall methodology with the proposed research questions. Then, we detail our experimental configurations.

4.1 Methodology

Following established work in question answering benchmarking [6], the approach that we use presents clearly delimited stages within

[1] Equinor Newsroom: https://www.equinor.com/news

[2] Llama2 training cut-off was July 2023 https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md#training-data. GPT4 training cut-off was April 2023 https://community.openai.com/t/what-is-the-actual-cutoff-date-for-gpt-4/.

[3] https://github.com/dariogarigliotti/ecai-pais-2024-EquinorQA

Table 1: Examples of questions and answers per question type in EquinorQA.

Question type	Context	Question	Answer
REG	'Equinor envisages a renewables portfolio that combines generation assets such as wind and solar with flexible assets such as batteries to help mitigate the intermittency of the renewable power generation. In the UK, we have our largest offshore wind power position as a company [...]'	What is an example of a so-called flexible asset that Equinor considers as part of renewable energy?	Batteries
MNF	The British Energy Security Strategy sets out an ambition for 95% of the UK's electricity to be low carbon by 2030, and battery storage systems can play an important role in this transition. They can store excess power generated from wind and solar and release it when the electricity grid needs the power most, improving security [...]	What percentage of the electricity in UK is expected to be low carbon by 2030?	95%
HARD	[...] In the UK, we have our largest offshore wind power position as a company with several offshore wind farms in operation and under development. In parallel, we are building our battery storage capacity, with our first asset Blandford Road in operation and our second asset Welkin Mill under construction [...]	What is the location of Equinor's battery storage capacity in UK, that is not Blandford Road?	Welkin Mill
	The COSLPromoter is already contracted to Equinor and will commence on the new contract in the first quarter of 2025. The firm contract is for one year, with options for a further four years. The COSLInnovator is contracted for two years, starting in the second quarter of 2025, and the contract includes options for a further three years.	Which firm has the longest initial contract term with Equinor: COSLPromoter or COSLInnovator?	COSL Innovator

the RAG paradigm. Specifically, first, a retrieval component obtains ranked results for each question from the passage collection indexed for search. Then, an augmentation stage aggregates the question-answer pair from each test instance together with the retrieved results, all within a well-engineered prompt that also captures the criteria to be required to an LLM. The final component performs the LLM-based generation of the supported answer for each question. On each component, we experiment with corresponding parameters of interest. We refer with '(question answering) method' to each instantiation of this RAG-based approach in a particular configuration, among all the studied parameters across all the components.

We conduct experimentation over EquinorQA with an ensemble of methods, in order to answer the following research questions.

- **RQ1**: How do Large Language Models perform for Question Answering over proprietary data?
- **RQ2**: How do parameter settings in the retrieval component affect the RAG performance?
- **RQ3**: What is the impact of different augmentation strategies?
- **RQ4**: How does RAG perform for the generation configuration?
- **RQ5**: How do our methods perform over data available during an LLM's training phase?

4.2 Experimental setup

Retrieval component. We built an index from the collection of uniquely identified passages, and retrieve the top 10 results for every question in EquinorQA with each of both methods, traditional lexical matching —lexical, for short— and learned dense retrieval —dense, for short—. Further sub-sequences of top 3 and top 5 results are obtained from these top 10 to experiment with smaller context sets. We perform retrieval with the established Pyserini library[4] as it provides off-the-shelf utilities to index and retrieve using the two methods.

Augmentation component. Through prompt engineering, we design a prompt that requests the LLM to produce the answer with the citation support in the desired format, with explicit requirements about (i) giving a brief answer and only if knowing it, and (ii) always citing and doing it only from the provided passages. The actual prompt template used in our experiments is shown in Table 2. We further aim to make observations on the phenomenon of a LLM possibly citing correctly most likely due to learning the pattern about the passages in the prompt being listed in the same order as the retrieved ranking, with the top result first. We experiment with alternative orders for the contexts: randomly ordered, or randomly but ensuring that the top result appears last.

[4] https://github.com/castorini/pyserini

Table 2: Template to build the prompt during augmentation.

Prompt template	Prompt template (ctd.)
You are an assistant for question-answering tasks. Use the pieces of context provided by the user to ANSWER the QUESTION to the best of your ability. If you don't know the answer, just say that you don't know. Keep the answer concise. Always cite one or more corresponding context IDs as your sources (which must be among the given CONTEXTS) between square brackets (e.g. [a1b2x34d]), as it's done in each example. Examples are given below, each example between the '⟨example⟩' and '⟨/example⟩' tags. After that, you are given the actual question with contexts so that you answer it.	⟨example⟩ ... ⟨/example⟩ ... QUESTION: ... CONTEXTS: Context ID: ... Context: ... ... ANSWER:

Generation component. For the final RAG stage, we generate supported answers by prompting established LLMs. Specifically, we use a family of open LLMs such as Llama2 [20] and two prominent models of the GPT platform, GPT3.5 (gpt-3.5-turbo-0125) and GPT4 [16]. Generation with Llama2 is performed by inference with HuggingFace transformers library, while for GPT models we access via OpenAI API.

This list is a summary of our experimental parameters:

- (Retrieval) Method: lexical or dense.
- (Retrieval) Cut-off: top 3, 5, or 10 results.
- (Augmentation) Number of examples: 0, 1 or 2.
- (Augmentation) Order of passages: as in ranking, random, or with the top retrieved result last.
- (Generation) LLM: open (Llama2-7b, Llama2-13b, Llama2-7b-chat, Llama2-13b-chat) or closed (GPT3.5, GPT4).
- (Generation) Maximum amount of new output tokens: 16, 32, 64.

Evaluation metrics. We evaluate answer correctness by verifying whether any of the collected possible expressions of the correct answer is an exact sub-string of the generation —exact match recall or EM recall, following [19]—. Since there is only one correct answer, we refer to it as answer accuracy. Answer support is evaluated by applying standard retrieval metrics of precision and recall with respect to the retrieved passage set (all the cited passage identifiers in the generation) and the relevant passage set (the set of all the known relevant passages such that they appear among the contexts provided in the prompt). For a given method, we report the average performance across all the questions of interest, whether is the full EquinorQA dataset or one of its distinguished subsets by question type.

Table 3: Experimental results for selected configurations, over specific sets of EquinorQA instances. In all these experiments, LLM is GPT4, and the prompt includes two examples. In each block, the best performance on a metric is shown in **bold**.

EquinorQA instances: All 60 questions					
Retrieval method	Retr. cutoff	Passage order	Answer Acc.	Citation Precision	Citation Recall
Sparse	5	Ranking	0.95	0.8667	0.8417
		Random	0.9167	0.8083	0.8
	10	Ranking	**0.9667**	0.8583	0.8417
		Random	0.9	0.7917	0.7833
Dense	5	Ranking	0.9	0.825	0.8167
		Random	0.8667	0.8	0.7833
	10	Ranking	0.9333	0.85	0.8222
		Random	0.9167	**0.8833**	**0.8556**

EquinorQA instances: 20 REG questions					
Retrieval method	Retr. cutoff	Passage order	Answer Acc.	Citation Precision	Citation Recall
Sparse	5	Ranking	0.95	0.9	0.875
		Random	0.9	0.85	0.825
	10	Ranking	**1.0**	0.95	0.925
		Random	0.9	0.8	0.775
Dense	5	Ranking	0.9	0.9	0.875
		Random	0.9	0.85	0.825
	10	Ranking	0.95	**1.0**	**0.9667**
		Random	0.95	**1.0**	**0.9667**

EquinorQA instances: 20 MNF questions					
Retrieval method	Retr. cutoff	Passage order	Answer Acc.	Citation Precision	Citation Recall
Sparse	5	Ranking	**1.0**	0.8	0.75
		Random	**1.0**	0.7	0.675
	10	Ranking	**1.0**	0.85	0.8
		Random	**1.0**	0.8	0.775
Dense	5	Ranking	**1.0**	**0.9**	**0.9**
		Random	0.95	0.85	0.85
	10	Ranking	**1.0**	0.85	0.825
		Random	**1.0**	0.85	0.825

EquinorQA instances: 20 HARD questions					
Retrieval method	Retr. cutoff	Passage order	Answer Acc.	Citation Precision	Citation Recall
Sparse	5	Ranking	**0.9**	**0.9**	**0.9**
		Random	0.85	0.875	**0.9**
	10	Ranking	**0.9**	0.775	0.8
		Random	0.8	0.775	0.8
Dense	5	Ranking	0.8	0.675	0.675
		Random	0.75	0.7	0.675
	10	Ranking	0.85	0.7	0.675
		Random	0.8	0.8	0.775

5 Experimental Results

We now analyze the main results for our experimental configurations. Throughout this section, Tables 3, 4 and 5 present the results for distinctive configurations among the multiple experiments conducted. The complete results tables for all query groups and methods are available in our repository[5]

5.1 RQ1: LLMs for SQA over proprietary data

Overall, we verify that the solutions to be built on LLM-powered methods have a clear advantage to make use of massive training data and regimes to address SQA and many other related tasks.

When comparing a method across the different query types, we observe particular patterns with respect to a combination of some of the studied parameters. (1) For the worst performing LLM, Llama2-7b, the results are mixed. (2) For the rest of our methods within the Llama2 family, the performance in the MNF group is higher than that on REG for the methods where the retrieval cut-off is lower (3 and 5), but it gets in most cases inverted to favour REG for the methods with the highest cut-off of 10. (3) In the case of GPT3.5 and GPT4 (the last one shown in Table 3), most of the cases exhibit MNF outperforming the same corresponding method over REG in answer accuracy. Lastly, HARD is for all configurations the worst performing group.

Few MNF questions are slightly harder, such as "What is the solar plant in commercial production that Equinor partially owns in Brazil, other than Mendubim?", or "What type of energy is predominant in Eirin field, oil or gas?" (gas) in a passage with main segment "Recoverable reserves in the field are estimated at 27.6 million barrels of oil equivalent, most of which is gas", both answered correctly by all the GPT4 models. In the cases of MNF outperforming regular, easy questions, we hypothesize that indeed this can be a mechanism to argue for the presence of contaminated data in the training phase of these LLMs, yet we cannot discard that the LLM is sufficiently capable to correctly answer just from the provided contexts.

5.2 RQ2: Retrieval component

In general, both lexical and dense retrieval methods lead to similar RAG results yet in several cases the lexical ones beat the dense performances, as it is the case that several questions are very specific, and so provide sufficient information to make traditional BM25 retrieval outperform its dense counterpart.

The results comparing the influence of different cut-offs are mixed. In most cases, the more contexts the better. This is the phenomenon when evaluating the whole EquinorQA and each subset by question type for citation precision and recall. Yet, taking as example the methods with dense retrieval, the results using a cut-off of 10 passages outperform those with cut-off of 5 (in particular for the REG group where the benefits are clearly large), except for the MNF group where a larger cut-off is mostly detrimental.

5.3 RQ3: Augmentation component

Here, we made two main observations. First, the number of examples clearly helps as the zero-shot scenario is too challenging to properly answer with the correct support format. Secondly, the order of the passages in the prompt is also influential, since we indeed verify that there is a tendency to favour the top ranked result when it appears as the first provided context. However, there are cases, including among

[5] https://github.com/dariogarigliotti/ecai-pais-2024-EquinorQA

the best performing configurations, where a random order of passages results to be comparable or even more convenient than the order given by the retrieval ranking. This is observed in the citation evaluation of the best performing method for the full dataset and for the question group with highest resulting measurements, REG.

5.4 RQ4: Generation component

Results from the ablation of the generation component are presented in Table 4. Overall, in line with reported results in the literature [6], GPT-based methods are the best performing ones, with GPT4 the best of them all, across all metrics and for all question subsets in EquinorQA. GPT4 has, in particular, a longer training phase by 19 months compared to GPT3.5, and shorter by very few months than the Llama2 models. Both Llama2 models optimized for chat beat their vanilla counterparts, as they better produce short, focused answers like the ones required by our dataset.

Regarding the parameter about maximum length of newly generated tokens, we consistently observe on the Llama2 models that the performances with shorter maximum length exhibit very comparative performances while generating increasingly faster the shorter they are required to be, which will most likely have a substantial impact on the large scale of an industrial application powered by it as it is not constrained to perform expensive with large maximum lengths.

5.5 RQ5: LLMs for SQA over public data

As a final experiment, we aim to contrast the results on our test collection with those on a general-purpose QA dataset that was already publicly available within the LLMs' training cut-off.

QAMPARI [1] is a QA dataset built with Wikipedia as its corpus of reference. In this dataset, each question is to be answered by a list of entities that occur in different relevant passages. Following the benchmark by Gao et al. [6], we randomly select 60 questions from QAMPARI, and refer to these as our QAMPARI instances. Each instance meets the selection criterion of having at least one of its possibly multiple correct answers occurring in the top 3 ranked results obtained by dense retrieval over the collection of Wikipedia passages by the benchmark. Furthermore, these instances are partitioned in three groups of 20 each, with respect to the three types of questions studied in QAMPARI: (i) simple (SIM), where each answer is reachable by one single hop in the underlying knowledge graph (e.g. 'Louvre' for 'Which cultural organi- zation is located in Paris?'); (ii) intersection (INT) of two simple ones (e.g. 'Which competition was won by Manchester City and had Manchester City as a participant?'); and (iii) composed (COM), where each answer is reachable by two or more hops (e.g. 'What is the height of buildings located in Dubai?').

In our experiments, we use the results ranked by dense retrieval obtained in the benchmark. For the augmentation stage, we use the same prompt template although we slightly change it where it mentions the citation format, from '(e.g. [a1b2x34d])' to '(e.g. [wiki:56781234])'. We evaluate our methods as before, observing only those with dense-based retrieval and only GPT4 as LLM from the GPT family, and now measuring the recall of all the known correct answers for each question. The behaviour requested to the LLM at generation time is to use the pieces of context provided in the prompt to answer the question. Hence, when measuring answer recall, we consider as relevant (i.e. golden or correct) answers only the subset of all known correct ones that appear in any of the contexts provided in the prompt. We evaluate citation precision and recall with an analogous criterion for defining the set of relevant passages.

Table 4: Experimental results for selected configurations, over the full set of 60 questions in EquinorQA. In all these experiments, the contexts are ordered as in the retrieval ranking, and the prompt includes two examples. For a given metric, the best performance on each LLM is in **bold** and the best overall performance is underlined.

LLM	Retrieval method	Retr. cutoff	Answer Acc.	Citation Prec.	Citation Recall
Llama2-7b	Sparse	5	0.45	**0.45**	**0.425**
		10	0.4	0.2861	0.2833
	Dense	5	0.4833	0.3667	0.3583
		10	**0.55**	0.4	0.3722
Llama2-13b	Sparse	5	0.35	0.3333	0.325
		10	0.4833	0.4333	0.4083
	Dense	5	0.3333	0.2	0.2
		10	**0.5833**	**0.5333**	**0.5139**
Llama2-7b-ch	Sparse	5	**0.6333**	**0.8**	**0.7667**
		10	0.5833	0.6833	0.6583
	Dense	5	0.6167	0.7	0.6917
		10	0.55	0.6685	0.6639
Llama2-13b-ch	Sparse	5	0.6667	**0.7667**	**0.7333**
		10	0.5833	0.6667	0.6417
	Dense	5	**0.7**	0.6833	0.675
		10	0.5833	0.6167	0.5889
GPT-3.5	Sparse	5	**0.85**	**0.8167**	0.7917
		10	0.8167	0.7667	0.7417
	Dense	5	0.8333	0.8	**0.8**
		10	**0.85**	**0.8167**	**0.8**
GPT-4	Sparse	5	0.95	**0.8667**	**0.8417**
		10	**0.9667**	0.8583	**0.8417**
	Dense	5	0.9	0.825	0.8167
		10	0.9333	0.85	0.8222

Table 5 presents results for distinctive configurations among all the methods we experiment with over the QAMPARI instances, contrasting GPT4 with the best performing Llama2-based LLM. A first observation is that our results are comparable with those obtained for QAMPARI with very similar methods by the benchmark. GPT4 is the best LLM across all metrics, and for all the instance sets except the COM question type subset in one metric, in most cases outperforming by a large difference, some with very substantial improvements.

Moreover, we make the following observations regarding the results for answer recall, when comparing corresponding methods in terms of the context order. (a) The results for GPT-based methods are very similar for both order by ranking and random in most configurations, but (a.i) the one for COM questions using 5 passages in the augmented prompt exhibits a strong preference for the order as given by the retrieval ranking, while (a.ii) for SIM questions with 10 passages random order clearly performs best. Overall, (b) randomizing the order of the ranked passages during augmentation is mostly hurtful for Llama2-based methods, in some cases very substantially, while GPT4-based configurations are mostly robust to this and in some cases they perform better due to it, which shows that it does not necessarily rely on this artefact to select the correct citations.

We also observe the results when duplicating the retrieval cut-off from 5 to 10. In particular, (c) for most cases with a method based

Table 5: Experimental results for selected configurations, over specific sets of QAMPARI instances. In all these experiments, retrieval is dense, and the prompt includes two examples. In each block, the best performance on a metric is shown in **bold**.

QAMPARI instances: All 60 questions					
LLM	Retr. cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
Llama2-13b-ch	5	Ranking	0.3633	0.4467	0.6308
		Random	0.3172	0.4231	0.63
	10	Ranking	0.2622	0.1744	0.2294
		Random	0.1529	0.1119	0.1372
GPT-4	5	Ranking	**0.624**	0.6842	0.6714
		Random	0.5937	0.6833	**0.6753**
	10	Ranking	0.5931	0.6404	0.6115
		Random	0.5806	**0.7039**	0.6264

QAMPARI instances: 20 SIM questions					
LLM	Retr. cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
Llama2-13b-ch	5	Ranking	0.3042	0.4392	0.6525
		Random	0.3333	0.5225	0.7125
	10	Ranking	0.225	0.1783	0.275
		Random	0.075	0.0458	0.0417
GPT-4	5	Ranking	0.7708	0.8025	0.8225
		Random	**0.775**	**0.8275**	**0.8242**
	10	Ranking	0.7275	0.755	0.785
		Random	0.7	0.7583	0.7462

QAMPARI instances: 20 INT questions					
LLM	Retr. cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
Llama2-13b-ch	5	Ranking	0.4858	0.5158	0.6067
		Random	0.3933	0.4708	0.6692
	10	Ranking	0.335	0.2642	0.2383
		Random	0.2071	0.1625	0.1742
GPT-4	5	Ranking	0.6667	0.7	0.675
		Random	0.6667	**0.7875**	**0.7267**
	10	Ranking	**0.6839**	0.669	0.6496
		Random	**0.6839**	0.7333	0.6496

QAMPARI instances: 20 COM questions					
LLM	Retr. cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
Llama2-13b-ch	5	Ranking	0.3	0.385	**0.6333**
		Random	0.225	0.2758	0.5083
	10	Ranking	0.2267	0.0808	0.175
		Random	0.1767	0.1275	0.1958
GPT-4	5	Ranking	**0.4345**	0.55	0.5167
		Random	0.3395	0.435	0.475
	10	Ranking	0.3679	0.4972	0.4
		Random	0.3579	**0.62**	0.4833

on the Llama2 LLM, performance clearly worsens. This degradation is substantial in the evaluation of citation metrics for Llama2-based methods. Yet, (d) for GPT4-based configurations, the variations in terms of this cut-off duplication are small. These differences in most cases are still negative with duplication, yet in few cases not only it improves but by doing so it achieves the best performance in a metric across all the methods on a question-type block.

As a case study regarding (c) above, consider the COM question "Where did a First-Circuit Appeals Court Judge of the United States attend college?" and its three correct answers {"Harvard Law School", "Harvard College", "Harvard University"}. Using a method based on Llama2-13b-ch, when 5 relevant passages are given in the prompt in random order, only the first two answers are relevant, and both are correctly in the generated answer (i.e. acc. = 1). With 10 passages, only one of the three, the last one, is in the response (acc. = 0.33). In the same comparison from 5 to 10, citation precision improves as more passages are cited, but the citation recall worsens (from 1 to 0.75) which is also a frequent phenomenon alongside (c).

6 Industrial Impact

Business intelligence professionals working at Equinor have registered an increasing interest for solutions to a variety of knowledge-intensive tasks in the company, across several *personas*, this is, categories of employees with particular duties and habits, and hence specific needs for information services within the organization. The deployment of information access solutions based on ChatGPT in Equinor, up to the date of March 8, 2024, amount to a total of around one billion tokens processed since October 2023. This volume corresponds to roughly 100k queries per month, from 1.5k to 3k unique users per week. The main languages used in these requests are Norwegian (Bokmål), English, Brazilian Portuguese, and other Scandinavian languages. Across the different information access scenarios, the common, most requested feature is to be able to process their own selected documents within Equinor proprietary data framework.

Our assessment of the LLM capabilities to address supported question answering as a fundamental task closely relates to our compendium of aspects of Equinor personas in terms of the information needs and corresponding data for this and other tasks that build on top of it. Table 6 presents an overview of main identified personas and their respective information needs, wishes and goals. Table 7 complements the description of these scenarios by showing statistics and details on the kind of data each persona gets to work with.

- An *incident analyst*, for example, reviews and collates information from all the incidents that happen during work in Equinor. They do so both to report to the government and improve safety. This role is carried out by few yet very important workers with a low output volume of sensitive data. This persona deals with many input documents about incidents, in informal language where domain-specific jargon abounds, but must produce formal summaries in the form of reports.
- A *citizen developer*, instead, analyses data and has learned some programming skills to do their job better. This persona deals with many documents shortly processing each, looking for extracting small pieces of information such as named entities, used for statistics that can support data-driven improvements across the organization.
- Yet another persona, an *oil platform maintenance worker*, works in a fast-paced, hostile environment and repairs and changes out equipment on the oil platforms. Workers under this persona deal at the same time with many of both 'work order' documents informally describing requested tasks and governing documents formally specifying these tasks. Their work dynamics made their need for information

Table 6: Equinor personas: needs, wishes, and goals.

Persona	Needs to	Wishes	Goal
Incident analyst	Collate information about incidents that happen on the platforms to put them in a report for the government	There were tools to more easily analyze and find subsets of the documents that cover the incident types and other information they are looking for	Live updates of incident statistics
Citizen developer	Develop dashboards and small tools to show reports about their domain	There was someone they could ask and talk to about how to use the low code applications they use to develop the view they produce	Automated "guru"
Communication worker	Write new articles, speeches, and reply to requests from journalists	There was an easier way than to both know who to ask and ask the right person about a given topic	Easy access and good lookup of previously written position pieces and articles
Supply chain purchaser	Be updated on the orders they manage	There was an easier interface to their data than SAP or SQL	To be able to collect the information they need in an easier way
Oil platform maintenance worker	Ensure safety, prevent long shutdowns, manage documents to ensure safety	They had a better overview of the incidents; there was better control from management over what should be done	Timely information about what they should be aware
Subsurface analyst	Investigate huge data sets to find small pieces of information about the oil or other interest under the surface	The documents they depend on were easier to read and explore; the information in the documents was better structured	Direct access to verified data so that they can use them for calculation
Production engineer	Be aware of previous "treatments" to the well	More structure in their logs and better "findability"	To get all old data into a new system that gives more structure

Table 7: Equinor personas: data statistics and characteristics.

Persona	Expected volume	Language (level)	Document length and structure
Incident analyst	Many (200k+ incidents)	Input: informal with jargon; output: formal	Input: Running text with structured metadata; output: medium length summary
Citizen developer	Unknown	Often jargon-heavy, informal	Input: running text containing facts; output: short text, process run often
Communication worker	Multiple 1k-10k shared documents on many topics, mostly business units	Formal	Regular length
Supply chain purchaser	10-100k	Jargon-heavy, many abbreviations	Very short texts
Oil platform maintenance worker	Multiple 100k (work orders); 20k+ (governing documents)	Jargon-heavy, informal (work orders); formal (governing documents)	Low-detail (work orders); long, detailed (governing documents)
Subsurface analyst	Multiple 1m documents	Variety: formal, jargon-heavy; Norwegian, English	Long running text with mixed in tables and bullet items
Production engineer	100s of documents (per field)	Jargon heavy, yet usually well written; Norwegian	Medium length; clear sections, mixed with images and tables

in short, crucial time and of high quality to ensure safety.
Using minimalist proof-of-concept systems, we have carried out initial observations among several employees in Equinor, for duties within personas where self-supported QA is a core component (e.g. a communication worker interested in the particular stance of the company on a subject). Notably, it allows the entire process to be no longer dependent on manually consulting multiple colleagues.

Clearly across this space of information access scenarios, the implementation and deployment of state-of-the-art self-supported question answering capabilities would have a decisive impact to enable achieving solutions for tasks like platform incident handling.

7 Conclusion and Future Work

In this paper, we have presented an analysis on the capabilities of LLMs to address supported question answering over proprietary data. We have built EquinorQA, a novel test collection from corporate news data in the energy domain, and have used it to assess a series of methods within the Retrieval-Augmented Generation general approach. We confirm the high capability of commercial closed LLMs and make observations on different question types in our dataset. We further compare several experimental configurations for each RAG component. We believe that our observations can shed light on the space of information needs for a variety of personas in Equinor, the largest energy company in Norway, who would benefit from state-of-the-art question answering solutions for multiple knowledge-intensive tasks.

As a future direction, we plan to further study aspects of data contamination in LLMs, for example, by comparing with a family of methods mainly consisting in fine-tuning LLMs over EquinorQA. Another possible area of work corresponds to the deployment of solutions based on the methodology here studied, and the measurement of their performance on actual question answering scenarios for a variety of settings, and in particular, oriented to assess across the different personas described in the previous section. This evaluation would be conducted both intrinsically by obtaining further expert annotations upon collecting proprietary user data within the company, and extrinsically by the means of testing the usefulness of these solutions within broader tasks like information extraction and summarization for business intelligence. A third line of research is to study the automatic labeling of test instances by LLMs, by evaluating how correlated this technique would perform with that of human annotators. A fourth direction would investigate the expansion of EquinorQA's reach onto other phenomena, by complementing it with additional instance sets addressing, for example, a different language like Norwegian, a finer-grained grouping of questions by type, and questions with different cardinality criteria for answers and citations. Furthermore, we would aim to obtain annotations also over other kinds of document collections within the same energy domain, which may be more relevant to other personas described above.

Acknowledgements

This paper is part of NRF project 329745 "Machine Teaching For XAI". We thank the anonymous reviewers for their comments. This work was funded by ValGrai, CIPROM/2022/6 (FASSLOW) and IDIFEDER/2021/05 (CLUSTERIA) funded by Generalitat Valenciana and PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe".

References

[1] S. J. Amouyal, T. Wolfson, O. Rubin, O. Yoran, J. Herzig, and J. Berant. QAMPARI: An open-domain question answering benchmark for questions with many answers from multiple paragraphs, 2023.

[2] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. PaLM 2 Technical Report, 2023.

[3] A. Asai, M. Gardner, and H. Hajishirzi. Evidentiality-guided generation for knowledge-intensive NLP tasks. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.162. URL https://aclanthology.org/2022.naacl-main.162.

[4] B. Bohnet, V. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig, K. Hui, T. Kwiatkowski, J. Ma, J. Ni, T. Schuster, L. S. Saralegui, W. W. Cohen, M. Collins, D. Das, D. Metzler, S. Petrov, and K. Webster. Attributed question answering: Evaluation and modeling for attributed large language models. 2022. URL https://arxiv.org/abs/2212.08037.

[5] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge. What's in my big data?, 2024.

[6] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398.

[7] A. Jacovi, A. Caciularu, O. Goldman, and Y. Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308.

[8] G. Jaimovitch-López, C. Ferri, J. Hernández-Orallo, F. Martínez-Plumed, and M. J. Ramírez-Quintana. Can language models automate data wrangling? *Machine Learning*, 112(6):2053–2082, 2023.

[9] E. Kamalloo, A. Jafari, X. C. Zhang, N. Thakur, and J. J. Lin. Hagrid: A human-LLM collaborative dataset for generative information-seeking with attribution. *ArXiv*, abs/2307.16883, 2023. URL https://api.semanticscholar.org/CorpusID:260334522.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[11] N. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467.

[12] Y. Liu, B. Deb, M. Teruel, A. Halfaker, D. Radev, and A. H. Awadallah. On improving summarization factual consistency from natural language feedback. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.844. URL https://aclanthology.org/2023.acl-long.844.

[13] I. Magar and R. Schwartz. Data contamination: From memorization to exploitation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL https://aclanthology.org/2022.acl-short.18.

[14] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147, 2022. URL https://api.semanticscholar.org/CorpusID:247594830.

[15] OpenAI. GPT-4 Technical Report, 2024.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

[17] O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

[18] Q. Si, T. Wang, Z. Lin, X. Zhang, Y. Cao, and W. Wang. An empirical study of instruction-tuning large language models in Chinese. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4086–4107, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.269. URL https://aclanthology.org/2023.findings-emnlp.269.

[19] I. Stelmakh, Y. Luan, B. Dhingra, and M.-W. Chang. ASQA: Factoid questions meet long-form answers. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566.

[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

[21] X. Ye, R. Sun, S. Arik, and T. Pfister. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, 2024.

[22] W. Yu, M. Jiang, P. Clark, and A. Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *ArXiv*, abs/2305.14010, 2023. URL https://api.semanticscholar.org/CorpusID:258841172.

[23] X. Yue, B. Wang, Z. Chen, K. Zhang, Y. Su, and H. Sun. Automatic evaluation of attribution by large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL https://aclanthology.org/2023.findings-emnlp.307.

[24] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023.