ISSN 1870-4069

Information Retrieval Techniques for Question Answering based on Pre-Trained Language Models

Ángel Cadena[1,3], Francisco F. López-Ponce[2,3],
Gerardo Sierra[3], Jorge Lázaro[4],
Sergio-Luis Ojeda-Trueba[3]

[1] Universidad Nacional Autónoma de México,
Instituto de Investigaciones en
Matemáticas Aplicadas y Sistemas,
Mexico

[2] Universidad Nacional Autónoma de México,
Facultad de Ciencias,
Mexico

[3] Universidad Nacional Autónoma de México,
Insituto de Ingeniería,
Mexico

[4] Universidad Nacional Autónoma de México,
Escuela Nacional de Lenguas,
Lingüística y Traducción,
Mexico

francisco.lopez.ponce@ciencias.unam.mx

Abstract. This paper presents a comparative study between two prominent pre-trained language models, RoBERTa and GPT-3, focused on their performance in Question Answering (QA). Broker exams serve as a rigorous evaluation guide, with which we examine the effectiveness of these models in understanding complex questions based on newly presented information in the form of the 19 Code of Federal Regulations of the United States (19 CFR). Our findings reveal insights into the strengths and limitations of each model, shedding light on their suitability for specific QA applications in the finance and legal domain. RoBERTa offers a fast implementation of a QA model yet it struggles processing complex questions, whereas GPT-3 is able to answer efficiently a wide range of reason based questions.

Keywords: Question answering, large language models, broker exam.

1 Introduction

Question Answering (QA) is a branch of Natural Language Processing (NLP) that focuses on the development of a model capable of efficiently answering human generated questions based on certain available information.

There are two main approaches to this problem: extractive QA which focuses solely on retrieving the specific data from a text and generative QA that creates text based on the information most relevant to the query. With the development of the Transformer architecture QA has evolved from fully controlled descriptive systems to pre-trained models fine-tuned to this task. Task focused datasets such as the Stanford Question Answering Datasets (SQuAD) version 1.0 [9] and version 2.0 [8], or the MLQA [5] a 7 language QA dataset, offer benchmarks for these models.

Extractive models like BERT [3] achieved an 87.4% of exact matches (EM), and 93.2 F1-Score on SQuAD v1.0, whereas in the English section of MLQA achieved 67.4% EM and 80.2 F1. A multilingual counterpart like XLM-RoBERTa [2] fine-tuned on SQuAD outperforms BERT in the English section of MLQA obtaining 67.8% EM and 80.6 F1. On the other hand, generative models like GPT-3 [1] outperform both of their predecessors, obtaining a 90.7% EM and a 93.0 F1 on the more complex SQuAD v2.0. In this paper we compare the performance and the results of an encoder such as RoBERTa against a decoder such as GPT-3 in this context.

We will work with publicly available USA Customs laws as well as Broker exams, with which we will demonstrate the capabilities of each model in terms of Information Retrieval and QA based on different levels of reading comprehension and user generated queries. In section 2 we show the conceptual differences between encoders and decoders, with information regarding the models chosen for this experiments.

In section 3, we explain the use of customs related information as well as the computational handling of legal texts needed for each model. Section 4 cover practical implementations, results and evaluations, in section 5 we present analysis of the results and in section 6 we conclude the article and present future work.

2 Background

Pretrained language models can be differentiated into three types: encoders, encoders-decoders, or decoders, depending on their intended use as well as their training methods. Encoders such as BERT create representations of text data based on masked language modeling. This type of model predicts a hidden token given a sentence, having access to every token in that sentence.

Decoders and Generative models like GPT focus on predicting the next word in a sentence, hence they only have access to previous tokens and their training procedure works with self-supervised learning. In terms of adjusting a model to a particular NLP task, encoders need to be fine tuned with a dataset focused on the chosen task, whereas decoders can avoid this step and work on the task without updating the model's parameters.

QA focused datasets help the fine tuning process but still fall short when dealing with unseen information. To solve this problem we need to feed each model the precise information we need them to process. Considering the computational expenses of a robust fine tuning as well as the necessity of having an appropriate dataset, we aim to compare these two architectures in order to obtain a QA model independent of fine-tuning and an increased dataset. The two models chosen for this comparison are RoBERTa and GPT-3.

Table 1. Structure of the 19 CFR.

Depth	Content	Reference
Chapter	U.S. Customs and Border Protection	I
Part	Special Classes of Merchandise	12
Section	Release under bond, liquidated damages.	12.3
Subsection	Bond amount	12.3 (b)
Subindex and text	Three times the value of the merchandise as provided in 113.62(n)(1)	12.3 (b)(2)

2.1 RoBERTa

RoBERTa [6] is an improvement of BERT [3] based on four modifications: (1) Training the model longer, with bigger batches, and over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. BERT [3] was trained in two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

The aim of MLM involves that given an input sequence of known tokens with one randomly replaced with the special token [MASK], the model should predict the actual value of the masked token. Next Sequence Prediction is a binary classification loss for predicting whether two sentences follow each other in the original text. In modification (1) the authors consider five English-language corpora of varying sizes and domains totaling over 160GB of uncompressed text and batch sizing set to 8k and the training steps were increased to 100k, 300k and 500k.

Modifications (2) and (3) come from a series of experiments conducted by the authors where they decide to remove the NSP loss. Each input contains full sentences sampled contiguously from one or more documents, such that the total length is at most 512 tokens. Inputs may cross document boundaries. If the end of one document is reached sentences from the next document are sampled using a separator token between documents. A final modification is the dynamic masking. The original implementation of BERT performs masking once during data preprocessing which results in a single static mask. For RoBERTa was used an dynamic masking instead, where the masking pattern is generated every time a sequence is given to the model.

2.2 GPT-3

GPT-3 presented in [1], is an auto-regressive language model trained with in-context learning that aims to solve several NLP tasks without the need of fine tuning the model on a task specific dataset. Transformer language models have been shown to increase performance based on the amount of parameters used during training [4], leading to GPT-3 being trained with 175 billion parameters, a substantial amount compared to BERT's 110 million. GPT-3 is a decoder based language model that works by queries (also called prompts) given by an user.

The model processes the user's query and generates an adequate response, but the response varies based on the structure of the query. GPT-3 was trained using three types of queries: Zero-shot, where there's no example of the intended use of the query; One-shot, where the query shows just one example of the task; and Few-shot, where the query shows multiple examples of the task in hand. For our task we focus on testing GPT-3's Zero-shot behavior in the context of Question Answering, this means that user prompts given to the model will be solely the questions with which we aim to evaluate. The only adjustment made will be a system prompt that modifies GPT'3 behavior for all iterations.

3 Code of Federal Regulations

For the theoretical evaluation of the aforementioned models we decided to work with customs laws of the United States. Legal information presents a considerable amount of free, accessible, and structured data easily adaptable to a QA problem. In particular Customs laws in the United States are an adequate branch given the existence of Broker Exams with complex questions, deep theoretical context, and Answer Keys that serve as an evaluation guideline for brokers and, in our case for a pre-trained model.

Broker Exams are composed of 80 questions based on the contents of 5 different legal documents. These questions are divided into 15 sections based on their subject and the answers are presented in a multiple choice format. The Answer Key contains the correct answer for each question in the test and includes the name of the legal document and the section within needed to answer the corresponding question. More importantly these questions range from standard information extraction to problem solving based on an explicit example, making them a more complex frame of work since understanding the context is a necessity in order to answer correctly.

The most referenced document, with more than half of the questions referencing it, is the Code of Federal Regulations (CFR) in the Answer Key from the October 2022 exam. The CFR is a collection of laws about numerous subjects written and presented yearly by the US Government divided into 50 titles. Title 19, named Custom Duties [7], is the most referenced text in this Answer Key. There is available an XML version of each of its chapters. This make it the best option in terms of relevancy and computational performance. We will refer to this text as 19 CFR onward.

The 19 CFR is comprised of four chapters containing 600 parts, each part referring to a precise set of information, additionally each part is divided into sections that contain a numbered subsection and, even though it doesn't always happen, a sub-index. Table 1 describes the levels of depth with an example of Part 12, Section 12.3, Subsection 12.3 (b), Subindex 12.3 (b)(2), as well as the corresponding text. We worked only with Chapter I, which encompasses parts 1-199, since those parts are the only ones mentioned in the Answer Key.

We filtered out the answers that do not come from the 19 CFR as well as those which require calculations or were questions of exclusion, leaving 44 questions adequate for an information retrieval task. After further analysis of the Answer Key, we observed that almost 40% of the remaining questions can be answered using only sections 111 and 113, thus we focused on a qualitative evaluation of those 16 final questions.

Table 2. Answers and prompts.

System Prompt	GPT-3's Response
You answer questions of a Broker exam.	Food and Drug Administration and Department of Health and Human Services.
You answer questions of a Broker exam. Answer with the explicit section used to answer the question.	Food and Drug Administration and Department of Health and Human Services. Section: 19 CFR Section (a)
Return the name of the section used to answer this question.	Animal and Plant Health Inspection Service and the Food and Drug Administration. Section used to answer the question: 19 CFR Section (a) Agencies within the Department of Homeland Security.
You locate sections of a legal document given a question.	The answer to the question cannot be found in the provided text.

4 Experiments

The first section describes the different type of questions used to evaluate, the following two sections detail the adjustments made for each model, containing technical details of each implementation, as well as the corresponding workflow. Results are reported at the end of the section. Prior to the use of any model sections 111, 113, and the filtered questions were extracted from both the 19 CFR's XML version and the October Broker exam, and finally adjusted into a text file. No labeling was done since the original file contains a basic label for each section, subsection and sub-index. There was no preprocessing applied.

4.1 Questions

Broker exams work incredibly well for question answering systems since the U. S. Customs and Border Protection uses a guideline in order to asses the validity of each question presented [10]. Said guideline states that exam worthy questions should "be developed and answerable from a designated exam source" meaning the context provided in the books should be enough, and that they should "reflect real-world situations that a broker might encounter", favoring clear questions over ambiguous ones as well as those that require deduction only from the facts presented in the question.

These conditions generate questions without exaggerated vocabulary, that are in theory simple to answer, are complete with the facts required to answer, and most importantly vary in their levels of complexity. This last characteristic is vital since it deepens the model's need for reading comprehension capabilities. We will observe each models performance on different type of questions in the result section.

Knowing that RoBERTa and GPT-3 vary in a theoretical approach and technique, we need to evaluate them using a comparable procedure. In order to do that, both models will be asked each filtered exam question without any modification to them. The context input for both is the corresponding section required to answer the question.

Table 3. RoBERTa's answers.

Q#	RoBERTa's answer	Exam answer
66	5 years	5 years
69	§\u2009111.96	Respond to the summons by providing CBP with the entry files requested because the broker is legally obligated to make such records available for examination by CBP.
75	an individual, partnership, association, or corporate broker	An existing permit holder.

4.2 RoBERTa

With access to HuggingFace's fine-tuned models specialized in Question Answering we avoided the process of fine-tuning and restrain our workflow to text preprocessing followed by the question based evaluation. We worked with a fine-tuned version of RoBERTa on the SQuAD2.0 [8] dataset. HuggingFace offers a simple procedure to perform Question Answering, we only need the chosen model, the sections that will serve as context, and the questions in a simple text form

Once readied the information with the use of a pipeline, we tokenized questions and context using the model's byte level BPE algorithm, we vectorized both elements and carried out the search in an efficient and simple way. RoBERTa answers with the most optimal response found in the text, meaning that even if the information retrieval procedure was correct we still need to check with the answer key since the answer might not be identical to the available ones in the multiple choice section, more about this results in 4.4.

4.3 GPT-3

In this case the workflow is more complex due to GPT-3 being able to perform tasks without the need of fine-tuning it first[1]. Nevertheless, we require additional procedure to be done so that the model is capable of answering based on information not presented during training. We worked with OpenAI's API and paid a fee to use GPT-3 unrestricted.

We followed two step workflow: the creation of embeddings for our datasets, and a search and evaluation over the embeddings using the corresponding questions. This strategy is optimal over fine tuning due to the fact that fine tuning adjusts the model's response to a specific type of prompt and differs from the Zero-shot learning approach. Similarly fine tuning doesn't alter the information the model has access to, making it unreliable for a QA task.

In order to create the embeddings, we used OpenAI's cl100k_base tokenizer without any preprocessing. Once tokenized, the embeddings are created using the text-embedding-ada-002 model that produces 1536-dimensional vectors. For further sections we chose the GPT-3 model gpt-3.5-turbo. GPT's responses are heavily reliant on the prompts given to the model. When working directly with the API, we have additional influence having access to a system prompt in addition to the user prompt.

Table 4. GPT-3's Answers.

Q#	GPT-3's answer	Correct answer
66	The broker must retain the records for at least 5 years after the date of entry or final withdrawal for a warehouse entry.	5 years.
69	The broker must comply with the Customs summons and provide CBP with the entry files for the ten entries prepared on behalf of Company Z within the last three years.	Respond to the summons by providing CBP with the entry files requested because the broker is legally obligated to make such records available for examination by CBP.
75	All licensed individuals, corporations, partnerships, and associations.	An existing permit holder.

In table 2 we show an example of the varying answers depending on the system prompt used, the user prompt in this example is Q70: A broker filed an entry for an importer of Irish tea. In addition to retaining the Customs Documents required to make entry and file entry summary for tea, which partner government agencies' forms must also be filed and retained? The system prompt is depicted in the table.

Upon verifying these responses with the answer sheet, we observe that the first three prompts provide the correct answer, yet the second and third indicate an incorrect section in the exam. The fourth one gives an incorrect answer. We established our user prompts to be the Broker exam questions without any modifications, whereas the system prompts evaluated will be the following 2:

- **Prompt 1:** You answer questions of a broker exam.
- **Prompt 2:** You answer questions of a broker exam. Answer with the explicit section used to answer the question.

4.4 Results

Questions in Broker exams are not always answered by purely extracting text from the corresponding section. The referenced information and the logic behind the question are fundamental to answer adequately. The most important task in our experiments is the analysis of each model's response to these varying circumstances. The following questions taken out of the evaluation set serve as an example of this phenomenon.

- **Q66:** Generally, how many years after the date of entry or final withdrawal for a warehouse entry must the broker retain the records?
- **Q69:** A duly licensed customs broker was served and named in a Customs summons signed by the Director of the Consumer Products and Mass Merchandising Center. The summons requires the broker to provide CBP with the entry files for ten entries prepared on behalf of Company Z within the last three years. The broker terminated Company Z as a client one year ago.

Table 5. Exact matches.

Model	Exact matches
RoBERTa	0.31
GPT-3 Prompt 1	**0.59**
GPT-3 Prompt 2	0.5

Table 6. Exam performance.

Model	Accuracy
RoBERTa	0.19
GPT-3 Prompt 1	**0.69**
GPT-3 Prompt 2	0.5

Company Z has not provided the broker with any specific written instructions regarding responding to a summons within its now revoked power of attorney document. How must the broker respond to the Customs summons?

– **Q75:** Who listed below is responsible for paying the annual user fee detailed in Part 111 of the Customs Regulations? A) Importers who file their own entries. B) An existing permit holder. C) All licensed individuals, corporations, partnerships, and associations. D) A permit holder reporting monthly employee new hires and terminations. E) A licensed entity with an employee embedded at a client's facility.

We see that Q66 is similar to the vast majority of those in QA datasets asking a straight forward question without any additional context, Q69, on the other hand, poses a fictional scenario the reader must analyze in order to obtain the correct answer, finally as a middle ground Q75 tasks the reader to review possible answers before actually answering. For further reference we will divide the questions in two sections: traditional QA questions, reasoning based questions. Before moving onto the final evaluations, we expose in detail the way our models RoBERTa answers these questions in table 3.

Upon inspection, we first see that Q69's answer is far from a legible piece of text and that only Q66's answer matches the correct response. Nonetheless, if we search for the specific section of Q75's answer, we observe an extract from Section 111.96 (c). We realized that that is the correct section from which the answer is extracted, verified with the answer sheet. This suggests that even if RoBERTa answers poorly, the model's semantic search should, in theory, be capable to adequately match questions with sections.

On the other hand, GPT-3 presents a considerably improved answer given the model's text generation capabilities. Fiddling with different prompts we were able to obtain answers anywhere from single sentences to multiple paragraphs of length. The inconvenience presented by is that the answer does not always match the precise wording of the answers presented by the exam, even if the answer itself is fundamentally correct, as shown in table 4 with Q69. In a similar sense tracing the text used to generate the answer can not be done as directly as with RoBERTa, this can be seen in table 2 as GPT-3 returns a nonexisten section of the 19 CFR.

Higher complexity answers implie that we must evaluate these models using more than just the exact matches metric. We evaluated them using both the exact matches metric as well as the correct answer for the corresponding question determined by the writer's criteria. In order to determine each model's test performance, two writers independently reviewed the answer returned by the models and cross referenced them with the corresponding answer sheet. Tables 5 and 6 show the evaluation.

Table 7. Extract of the evaluation.

Question	RoBERTa	GPT-3 (Prompt 1)
3: What type of bond is needed to operate as a custodian of bonded merchandise?	Single transaction bond.	Basic Custodial Bond.
73: Under what circumstances may a licensed broker accept fees from an attorney?	The charges are not paid bythe broker.	A licensed broker may accept fees from an attorney if the amount of the fee is commensurate with the time, effort, and skill expended by the broker in performing his services.
76: What must a corporate broker do to continue to conduct Customs business after thecorporate office who qualified its license retires?	Provide written notice.	A corporate broker must provide written notice to the Assistant Commissioner when the corporate officer who qualified its license retires. The broker must also send a copy of the written notice to the director of each port through which a permit has been granted to the broker.

To further discuss the results we present table 7 that shows 3 questions selected by their different complexity levels as well as each model's answers. The analysis of this table is done in the following section.

5 Analysis and Conclusion

RoBERTa is simple to implement and free option to work with. Nonetheless this model, mainly due to its encoder nature, lacks the capability to answer more complex questions as the ones shown in this paper. The questions used for evaluation ask more of the model rather than just locating the point in the text in which the answer is contained. Reasoning based on information is required in order to answer correctly questions such as Q69 and Q75, which tend to be the most frequent type of of question in the exam. GPT-3 as a decoder outperforms considerably its encoder counterpart, again, due to the nature of the exam questions.

This result suggests that language models that are able to analyze context and answer based on certain information are bound to perform better in complex scenarios of question answering. However GPT-3's prompt depending answers make it susceptible of a lower performance as seen in table 6. A more complex prompt, even if just by a sentence, endangers the model's performance on the test Table 7 exemplifies both model's behavior and limitations for this task.

For question 3 we see that both models manage to respond with a type of bond, even if RoBERTa's answer is incorrect (the answer is Bond Type 2) the model manages to return an clear answer to the question. GPT-3 answers correctly even though the exact match differs, a simple online search of what are Bond Type 2 returns Basic Custodial Bonds. Q73 provides an example of a question that RoBERTa can't answer correctly, its response is not related to the question. Remarkably GPT-3 answers almost exactly as the broker exam needs, this discrepancy might be due to the phrasing of the question itself.

Datasets like SQuAD contain mainly W and H questions[1] without modifiers e.g., What was the name of Beyoncé's second solo album? or What do greenhouses do with solar energy?. Even if broker exam questions are not ambiguous, their semantic properties make them more complex that the average QA question, hence decoder models answer them in a more efficient manner. Finally Q76 shows us that generative models are not optimal and that users should be critical with their answers.

GPT-3's response seems reasonable and well structured yet the correct answer differs substantially: Appoint a new broker as an officer of the corporation and notify CBP of the new license qualifier. Summarizing, decoder and generative models seem to be a better option for dealing with complex question answering scenarios. For this particular task even though RoBERTa's implementation was free and simple its performance overshadows these aspects. On the other hand even though GPT-3 might need a more complex workflow as well as a monetary involvement, its results are worth the effort. With minimal user prompt engineering (zero-shot scenario) the model achieved acceptable results.

6 Future Work

Future work requires a deeper and better structured dataset that contains all of the 19 CFR, not only specific sections, as well as the other four mentioned books that serve as a guide for Broker exams. Similarly a precise separation of subsections, as well as tagging might prove to be beneficial for section extraction with models like GPT-3. A long term project would be the creation of a dataset using complex questions specialized to extractive or generative models in order to carry out a fine tuning. Additionally, regarding the evaluation methodology, these experiments would benefit from verifying the answers with legal experts, brokers in particularly, in order to account for linguistic variations presented by generative models.

Acknowledgments. We thank CONAHCYT (CF-2023-G-64) and PAPIIT (IT100822) for the financial support for this paper.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., et al.: Language

[1] Who? What? When? Where? Why? How?

models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (2020) doi: 10.48550/arXiv.2005.14165
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451 (2020) doi: 10.18653/v1/2020.acl-main.747
3. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019) doi: 10.48550/arXiv.1810.04805
4. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling Laws for Neural Language Models. arXiv (2020) doi: 10.48550/arXiv.2001.08361
5. Lewis, P., Oguz, B., Rinott, R., Riedel, S., Schwenk, H.: MLQA: Evaluating cross-lingual extractive question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7315–7330 (2019) doi: 10.18653/v1/2020.acl-main.653
6. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. In: Proceedings of the International Conference on Learning Representations, Conference Blind Submission, pp. 1–15 (2019) doi: 10.48550/arXiv.1907.11692
7. National Archives Records Administration's Office of the Federal Register and Government Publishing Office: Title 19 of the code of federal regulations. GovInfo (2022) www.govinfo.gov/app/collection/cfr/2022/title19/chapterI
8. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 784–789 (2018) doi: 10.18653/v1/p18-2124
9. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100000+ questions for machine comprehension of text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392 (2016) doi: 10.18653/v1/d16-1264
10. U.S. Customs and Border Protection: Customs broker examination guidelines for writing new questions (2018) www.cbp.gov/trade/programs-administration/customs-brokers/license-examination-notice-examination/guidelines-writing-questions