Introducing CQuAE : a New French Contextualised Question-Answering Corpus for the Education Domain

Thomas Gerald[1], **Louis Tamames**[2], **Sofiane Ettayeb**[2], **Patrick Paroubek**[1], **Anne Vilnat**[1]
[1]Université Paris-Saclay, CNRS, LISN
{firstname.lastname}@lisn.upsaclay.fr
[2]Stellia
{firstname.lastname}@stellia.ai

Abstract

We present a new question-answering corpus in French designed for the educational domain. To be useful in such a domain, we have to propose more complex questions and be able to justify the answers on validated material. We analyze some properties of this corpus. The last part of this paper is devoted to the presentation of the first experiments we carried out to demonstrate the value of this dataset for learning a Retrieval Augmented Generation framework. Different experiments are proposed, with an automatic evaluation. A human evaluation is finally exposed to confirm or infirm this automatic evaluation.

Keywords: Resources, Question-Answering, Education

1. Introduction

This work takes place in the important area of education, particularly with the objective to help students learn or revise their lessons by providing them with exercises made up of questions and associated answers. To ensure that both questions and answers are in line with the course, teachers should provide large sets of exercises of this type, basing the questions and of course the answers on their lectures or duly validated. The aim is to go beyond the simple factual questions that are easy to answer and to be able to ask complex questions that go beyond the search for a named entity answer. For example, faced with a course on the beginnings of the French Revolution, we do not want to only ask when the storming of the Bastille took place, but also what the reasons led the demonstrators to this outcome, thus coming closer to the course questions a teacher might ask. To help teachers in this process, which would be intensely time-consuming for them, we are working on the automatic constitution of such a corpus, starting with the manual construction of an initial one. At present, no corpus meets all these criteria, i.e. questions and answers that can be complex, that are based on validated but short sets of documents (the teacher's lessons), and, are in the French language. We would like to work on several disciplines, and possibly at different teaching levels, but we've started our study with history as taught at the end of middle school and the beginning of high school. To have a basis for comparison, we also carried out a few tests on Geography, Life Science, and Civil Education. We thus built up a corpus containing :

- questions created from course documents, not only factual questions but also broader, more complex ones
- answers that are either extracted from the course or constructed from several elements scattered throughout the document,
- the source document, which validates both the interest of the question and the quality of the answer produced.

With those different elements, this corpus could be used to train components of a Retrieval Augmented Generation (RAG) framework where retriever and generator are needed. To this end, we propose in this work to measure the adequacy of the dataset to develop such application, evaluating the different components of the framework.

We will first detail how we collected a new corpus designed for the educational domain. Then, we present an analysis of this corpus to introduce its content. The last part of this paper will be devoted to present experiments we have carried out to demonstrate the value of this dataset for learning a RAG framework and how we can automatically produce answers based on reliable materials, as required in the educational field.

2. Related Works

Automatic summarization, questions, and answers generation have been and remain central topics in the NLP community. These different tasks benefited from machine learning and deep learning advances. The "transformer" neural architecture proposed by Vaswani et al. (2017) has provided significant improvements for generation. These architectures have been revised in many ways by addressing multi-tasks as in Raffel et al. (2020) or (Radford et al., 2019) or by increasing the size

LREC-COLING 2024, pages 9234–9244

of the models and datasets as in Brown et al. (2020). Primarily developed for the English language these pre-trained models are now available in French with CamemBERT and FlauBERT (Martin et al., 2020; Le et al., 2020) language models (LM) or the BARThez generation model (Eddine et al., 2021). Most of the effective approaches now consider multi-lingual settings for pre-training LM (Liu et al., 2020).

Adaptation. To adapt these language models to a specific task, a common approach consists of fine-tuning the model on new data. However, this process is prohibitive in terms of processing time and capacity, particularly on larger models. To this end, different approaches have emerged particularly for large models, such as adapter (Pfeiffer et al., 2020) consisting of fine-tuning only some additional layers added between original neural blocks. Closely related to the adapter approach, the LoRA (Hu et al., 2022) methods have been proposed this year, this methods rely on adding the results of a 2-layer perceptron to the linear functions defined in the model; one of its advantage is that it does not need supplementary computation (factorizing before inference) for efficiently adapting to a new task.

Corpus. The corpus SQuAD (Rajpurkar et al., 2016) strongly participates in improving question-answering models, providing a large dataset of questions and extractive answers. More recently, Google published the corpus Natural Question (Kwiatkowski et al., 2019): a corpus with natural language questions, with long and short paragraphs for answers (extracted from the English Wikipedia). In conversational QA the corpus CANARD and QUAC (Elgohary et al., 2019; Choi et al., 2018) are available. For retrieval-based question-answering where documents are answers, the MS-Marco passage dataset (Nguyen et al., 2016) is today the reference for training or fine-tuning models. If most QA corpora are available in English, the French community also produced corpora such as FQuAD (d'Hoffschmidt et al., 2020), Piaf (Keraron et al., 2020) or CALOR-QUEST (Béchet et al., 2019) for extractive QA. More recently, the CALOR-DIAL (Béchet et al., 2022) corpus addresses dialogue question-answering for the French language. However, these corpora mainly rely on factual QA, where the answer is a short text such as a named entity, an event, a date, a quantity, or a location. Recently, a new corpus Autogestion (Antoine et al., 2022) has been created to address non-factual questions, the associated study demonstrates the inability of standard models to address most complex questions.

Prompt tuning and LLM. Recent works have focused on explainable answers by Chain of Thought prompting (Wei et al., 2022) leveraging huge language models. Similarly (Huang et al., 2023) proposed improvements to these approaches with no additional data needed. For those large models pre-trained on a huge amount of text, a huge quantity of data is not necessarily required to adapt to a specific task. For instance, the LLaMA model (Touvron et al., 2023) is effective when fine-tuned with highly qualitative data. In (Zhou et al., 2023), the authors showed the importance of the data quality over the quantity, which encourage us to leverage our dataset on the fine-tuning of an LLM. Lately, models like Mistral-7b (Jiang et al., 2023) showed better performances than twice bigger models such as llama2-13b, and showed that smaller LLM is a better choice in many use-cases. Recent experiments have been proposed using the open.ai ChatGPT solution to generate instruction-based data and train models on it (Taori et al., 2023; Zheng et al., 2023) producing positive results for complex generation.

3. Collecting CQuAE

To gather a qualitative French corpus for education, we collected schoolbooks content from middle and high school mostly about History, Geography, but also biology, and Civic Education from the "Livre scolaire"[1]. In addition, we retrieve Wikipedia articles related to education. We filter them using Wikipedia API with queries based on the titles from the educational textbooks, then we bring together the subsections selected. To avoid having too large contents to read, we decompose Wikipedia articles considering at most three paragraphs for a document (within the same section). Thus a Wikipedia article corresponds to many documents in our corpus. We collected 3 891 documents (only 1 122 have annotations), composed of 14 433 paragraphs (3 893 with annotations).

To collect the corpus, we present a paragraph to the annotators and ask them to create the following annotations:(a) **a question** written by the annotator; (b) **the question type** which may be factual, definition, course or synthesis; (c) **the question support(s)**, i.e. extracted spans targeting the subject of the question;(d) **answer element(s)** i.e. the different passages allowing to answer the question and (e) **the handwritten answer** from the annotator, using the answer elements.

It should be noted that, for each document, we ask annotators to create about 10 annotations (and more if possible). Table 1 gives examples of questions, and supports these questions.

[1] https://www.lelivrescolaire.fr/

Type	Question	Support
Factual	In which year did Christopher Columbus reach America ?	Christopher Columbus reached America (1492)
Definition	What is a rotary press ?	A rotary press is a typographic press mounted on a cylinder, allowing continuous printing.
Course	How did the Europeans legitimize their domination?	Europeans rethink the hierarchy of people within a Christian and European-centered scheme which then serves to legitimize their domination
	What are the names of those who indicate how to practise the Muslim religion? According to which text do they do this?	It is the ulemas who regulate religion on the basis of Sharia law.
Synthesis	Why did some French people support the state of emergency after the 2015 Paris attacks ?	• protects them against the terrorist threat and the risk of a new attack, which is feared by all. • This exceptional regime continues to appear as “a necessity”.
	Who needs to be involved to fight climate change according to Matt Petersen? How do we do it?	Matt Petersen works for the sustainable development of the city of Los Angeles, alongside the city’s mayor [...] we need everyone. All smiles, the mayor of Los Angeles has connected [...] solar panels installed on private roofs [...] ...
	Why does this article call the midinette movement a ”victory for feminism”?	Midinettes should not be disparaged. It is not in good spirit to tax them with frivolity because they work in dresses, they are young and pretty and [...] on of woman, exercised in these tragic...

Table 1: Examples for the four question types (translated from French)

One of the main objectives of our corpus is to deliver questions requiring different levels of expertise, depending on the type of information needed to answer them. This ”difficulty” level is related to the *question type* field which is one of the following:

- **Factual**: The answer is a fact or a list of facts (event, person, location, date...).
- **Definition**: The answer corresponds to a definition of a concept or a word.
- **Course**: The answer is not a fact or a description but contains explanations or many details. However, it must be explicit in the context.
- **Synthesis**: The answer relies on different elements of the text and different pieces of information must be gathered or it involves interpretation in order to answer the question.

To ensure we have enough complex questions and not only factual or definition ones, we explicitly asked annotators to follow a specific ratio, 40% of factual and definition, and 60% of synthesis and course questions. We also instruct annotators to avoid creating synthesis and course questions when the document does not offer the possibility to create them, this ratio is thus not strictly observed. Two annotator groups have been working on the dataset: the A group with no specific teaching backgrounds but educated; and the B group having knowledge and a specific educational background.

In Table 2 we report the current question type distribution obtained by both groups. Notice that, in proportion, the first group (A) produced more *course* questions than the second (B with educational background), while the B group produced more *synthesis* questions. Nevertheless, both distributions are quite similar with, for both of them, a preference for *course* questions.

In order to ensure the quality of the questions, we asked annotators to judge (and correct) other annotations, this score is not studied in this current work. For more details on the annotation procedure, we made available an annotation guideline[2]. The corpus is available on an anonymous github[3].

Qu. Type	Group A		Group B		Total	
Course	4 784	47	490	38	5 274	46
Factual	2 106	21	294	22	2 400	21
Synthesis	1 756	17	338	26	2 094	18
Definition	1 506	15	181	14	1 687	15
Total	10 152	89	1 303	11	11 455	100

Table 2: Question types (# and %) for each group (A=educated, B= with educational background)

4. Corpus Statistical Analysis

In this section, we propose to study the different properties of the corpus we collected. To assess the relevance of the different choices, we analyze different aspects starting with the study of the length of the questions and the answers. We then analyze some linguistic features.

4.1. Question and Answer length

To have a better idea of the questions produced, we compute their length, following the different question types. We compare these results to the two other French QA datasets mentioned earlier, i.e. FQuAD (d’Hoffschmidt et al., 2020), Piaf (Keraron et al., 2020), where all the question types are mixed. Figure 1 illustrates this study. We can observe that in our corpus, the *definition* questions are the shortest ones, with a small deviation. The three other types are rather similar, but the most difficult ones

[2]this guide will be given in the final version

[3]https://gitlab.lisn.upsaclay.fr/gerald/cquae

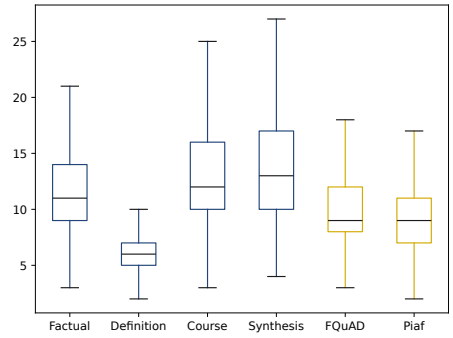

Figure 1: Question length depending of the question type in our corpus, compared to FQuAD and Piaf

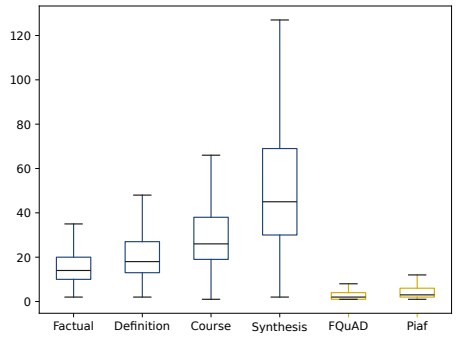

Figure 2: Answer length depending of the question type in our corpus, compared to FQuAD and Piaf

are the longest ones: the *synthesis* questions, followed by the *course* questions. We can observe that the deviation is rather important for all of them. FQuAD and Piaf are quite similar to each other, with shorter questions and smaller deviations. We can observe that the *factual* questions are the most similar to the questions of these two datasets. It is not surprising, as most questions in these datasets are factual.

The same study has been carried out on the length of the answers. We propose two answers: the part of the document in which the answer is found, and the answer composed by the annotator from this extract. In Figure 2 we see that the *synthesis* answers are clearly longer than the others, some of them being very long. The factual answers are the shortest. We can also see that the answers coming from FQuAD and Piaf are significantly shorter than the answers in our corpus, even in the case of *factual* answers. We thus obtain more often longer questions and always longer answers. It is not surprising for the complex questions (such as *course* and *synthesis*, but it is also the case for the other types of questions. We can conclude from this first study that the questions composed by the annotators are mostly longer than those generally found in the literature, even for the classical *factual* questions. This phenomenon is even more pronounced for the answer lengths. It is probably due to the educational context, where you cannot only give a simple answer, but you have to justify it. Thus, the classical existing datasets cannot give us good examples.

4.2. Linguistic features

After the length of the questions and answers, we study the words used to ask the different types of questions. We want to see if the interrogative forms are different, and also to compare them with each other. The following figures (Figure 3, 4 and 5) illustrate this point.

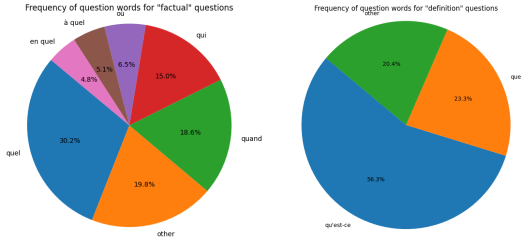

Figure 3: Interrogative word at the beginning of a *factual* question (on the left) and of a *definition* question (on the right)

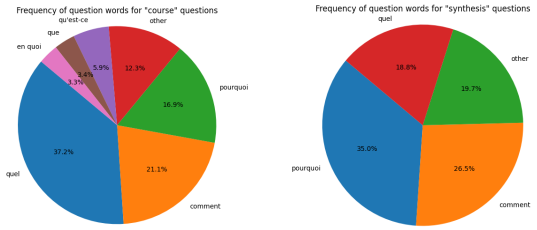

Figure 4: Interrogative word at the beginning of a *course* question (on the left) and of a *synthesis* question (on the right)

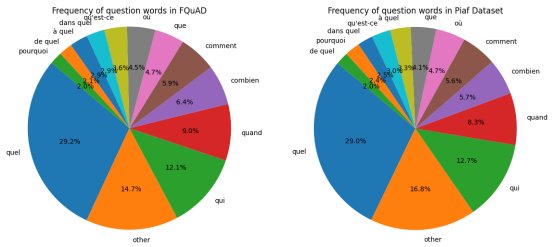

Figure 5: Interrogative word at the beginning of a FQuAD question (on the left) and of a Piaf question (on the right)

The distribution of interrogative words is quite different depending on the type of questions. For the *factual* ones, we find a majority of *which* or *what* in English (both corresponding to *quel* in French), and some variants as *in which* or *at which*, followed

by *when*, *who* and *where*. The *definition* questions are very specific, with quite only the word *what is* or *what* (in French *qu'est-ce* or *que*) to introduce the questions. The more complex questions make use of *why* and *how* in addition to *quel* in French (which may be *which* or *what*) in English, with a rather comparable distribution, even if slightly different. So, we can draw two conclusions: *definition* questions are quite different, and the complex questions are often introduced by *why* and *how* which are not used or very rarely in the other types of question. If we compare these results with the distribution of existing datasets as FQuAD and Piaf, as illustrated in Figure 5, it is not surprising to see that they are quite similar to each other, and also quite similar to the distribution of our *factual* questions: a lot of *what, who, when* and very few *why* or *how*.

To conclude this analysis, which could be further developed, we observe that, compared to the other well-known datasets in French, we have longer questions and really longer answers, and we have a large number of questions introduced by *why* and *how*, as we wanted to deal with complex questions in the educational domain.

5. Experimental Protocol

In this section, we describe the methodology and the choice made to develop the different parts of a Retrieval Augmented Generation (*RAG*) framework. We first describe the dataset, then the retrieval procedure (how to retrieve relevant documents associated with a question), and finally the settings to train the generator.

5.1. Dataset

To fine-tune the different components of the RAG framework, we split our corpus into a training, validation, and test set. We first gathered all questions belonging to the same document in order to avoid the model having the same documents during training and testing steps. We sampled different splits and selected the one where train and test sets have the most similar proportion of schoolbook documents, and consequently the same proportion of Wikipedia documents. We summarize in the table 3 the characteristics of the different splits.

	Train	Validation	Test
N-questions	10 490	407	558
definition	14%	14%	19%
factual	21%	22%	19%
course	46%	45%	45%
synthesis	18%	19%	17%
schoolbook	46%	54%	45%

Table 3: Characteristics of the splits, with N-questions the number of question in each set, and definition, factual, course, synthesis the percentage of each question type. The last row give the portion of schoolbook documents on total number of collected documents (wikipedia and schoolbook).

5.2. Document Retrieval

In order to generate a correct or adapted answer, we would base the generation process on a limited collection of verified documents. In order to generate accurate and contextually appropriate responses, our approach initiates the generation process by employing a retrieval step, which supply a limited set of verified documents. Thus it is natural to consider as the first step of the pipeline a retrieval process to yield an answer derived from relevant educational document passages. We can easily judge the relevance of the paragraphs retrieved, as the questions are created on the basis of a document where annotators have selected the location of the answer elements.

If a paragraph contains part of the text that has been selected by the annotators we would consider the paragraph as relevant to the question. Notice that we consider as paragraphs collection, all paragraphs collected whether or not it has been manually annotated (see section 3).

Models. If today a large number of models are available for English language retrieval, only a few are trained on the French language. For the experiments, we choose to evaluate the following approaches:

- BM25 (Robertson and Zaragoza, 2009): The classical retrieval baseline based on TF-IDF.
- DPR (Karpukhin et al., 2020) : A Dense Passage Retrieval approach, based on representation embedding and comparison between documents and questions. Notice that we considered as a pre-trained model a French version of BERT fine-tuned on PIAF and FQuAD corpus[4].
- DPR-FT: The same DPR model fine-tuned on our corpus using the Haystack framework[5] for fine-tuning and ranking (during two epochs)

For training the DPR-FT approach we considered hard negative paragraphs (wrong passages with relevant words), paragraphs retrieved from BM25 having a rank between 30 and 40.

[4] https://huggingface.co/etalab-ia/dpr-question_encoder-fr_qa-camembert

[5] https://haystack.deepset.ai/

5.3. Learning to generate questions and answers

One of the multiple uses of our corpus is to learn how to answer questions according to a source document without using external information, the goal is to limit hallucination and control the content shown to users. Here we focus on the Large Language Models we fine-tuned for this task. We choose the LLaMA2-7b and Mistral-7b models for their performances on many public benchmarks. We limit our experiments to 7b parameters models due to computational limitation, in future works, we might explore bigger model.

We fine-tuned the two algorithms on the task of question answering, given the instruction 'Réponds clairement à la question en te basant exclusivement sur les paragraphes associés.'[6], followed by the question and the associated paragraphs used by the annotators to write the answer. With these settings, we launch fine-tuning of those two models for 3 epochs on the dataset on an A100-80go. We use the Low Rank Adaptation method (LoRA), $bf16$ precision, and $4bit$ quantization. We set the maximum input size at 4096 tokens as our longer sample is around 3000 tokens. We used a batch size of 64 (using gradient accumulation). In the end, we choose the checkpoint with the lowest validation loss after 1.5 epochs.

In our study, we employed the Rouge unigram score (R-1), ROUGE-L (R-L) score, and the SacreBLEU score as our primary evaluation metrics. By focusing on recall, ROUGE-L ensures that the generated text contains important information from the reference. SacreBLEU compares n-grams in the reference and generated texts. We acknowledge that those metrics cannot entirely validate the relevance of the generated answers, so we also consider a human evaluation.

6. Analysis

6.1. Document Retrieval

In this section, we address the retrieval problem, the question is whether or not we can use ranking systems to retrieve the correct documents. We summarize the performances of the three ranking systems discussed in section 5.2 in the table 4. Paying attention to the BM25 performances we would first observe that $P@1$ reaches $.53$, which means that in half the cases the first document retrieved is one of the document selected by the annotators. As BM25 methods rely on bag-of-words, we can suppose that rare or specific tokens (or words) are present at least in half of the queries. In parallel, the Average Precision at 10 ($AP@10$) is slightly higher compared to $P@1$ but still close to each other. Therefore, we can advance three hypotheses to explain this behavior: remaining questions (about 40%) rely on complex linguistic structures where bag-of-words approaches struggle; remaining questions have missing context elements; many documents can answer the questions. To complete this study, we additionally provide experiments using language model approaches for ranking. However, the BM25 method performs better in all cases, with more relevant documents in the first position ($P@1$) and a better ranking of relevant documents ($nDCG@10$ and RR metrics take document position into account in their calculation). Indeed, ranking approaches based on complex transformer models rarely outperform feature-based approaches for the French language. This is particularly due to the lack of retrieval corpus in the French language, as there is no equivalent in French of a large corpus as MSMarco (English corpus). Even if offering poor performances, we should observe that fine-tuning the DPR model on our data always gives similar or better performances. This behavior underlines that our corpus has some intrinsic specificities that are not present in the FQuAD-like corpus.

Ranker	P@1	RR	nDCG@10	AP@10
BM25	**.53**	**.62**	**.67**	**.59**
DPR	.43	.52	.54	.50
DPR-FT	.43	.53	.56	.51

Table 4: Ranking performances on our corpus for the different approaches (see section 5.2)

While results depicted in the table 4 give information on how many original documents/paragraphs are retrieved, it does not inform if the retrieved document is relevant to the topic of the question. To verify whether the documents retrieved were out of topic or not, we designed a small experiment where human evaluators were asked to decide if the first document retrieved by the BM25 method is related or not to the question. Considering six annotators we asked them to decide on 20 questions each. On the evaluated set (120 questions), in 75.8% of the cases the documents retrieved were relevant to answer the question. This is much higher than the $P@1$ performances (53%), thus, we can state that in some cases, many paragraphs in the corpus contain an answer to a question.

6.2. The Retrieval Augmented Generation (RAG) framework

In this section, we will consider the whole RAG framework with a simple automatic pipeline with at the first stage the BM25 ranking approach (retrieval

[6] Give a clear answer to the question, basing your answer exclusively on the relevant paragraphs.

part) and at the second stage the question answering approach (generation part). We evaluated three configurations: the models without any adaption in a zero-shot setting (ZS), the models fine-tuned on our corpus (FT), and, the models fine-tuned using instead of documents provided by annotators the documents retrieved by the BM25 approach (FT-R). Notice that only the last configuration can be considered as a RAG framework. The results of the different configurations are reported in the table 5. In our fine-tuning protocol according to results,

Config	Model	R-1	R-L	BLEU
ZS	LLAMA2	.18	.14	4
	Mistral	**.34**	**.29**	**13**
FT	LLAMA2	**.52**	**.45**	**23**
	Mistral	.41	.35	14
FT-R	LLAMA2	**.47**	**.35**	**14**
	Mistral	.36	.30	11

Table 5: Scores obtained with the different configuration of Llama2-7b and Mistral-7b models.

it seems that all models benefit from fine-tuning whatever the metric considered. The Mistral model obtains better performances for the zero-shot configuration than the LLAMA2 model, in fine-tuning we get the opposite conclusion. However, while the Mistral model seems to have poor performances regarding the fine-tuned results, it is still difficult to state if it generates fewer correct answers. Indeed, ROUGE and BLEU compute a score based on the number of common n-grams between the text and the references (in our case only one reference), thus Mistral may eventually make extended use of synonymous or different sentence construction.

6.3. Influence of the question type

As described in the section 3, in addition to questions, answers, and relevant passages we asked annotators to select a category for each question (definition, factual, course, and synthesis). Those categories implicitly represent the difficulty level of each question. Thus, we propose to analyze results in terms of n-grams-based scores for each of the question types. We reported performances in the table 6 for both LLAMA and Mistral fine-tuned models. Looking at the different models' performances we can observe that for all different question types, the Mistral model got lower scores. It therefore tends to confirm the hypothesis that our fine-tuned version of Mistral tends to produce more incorrect answers. Particularly, Mistral got the lower score for factual questions while generally, the answer is often a named-entity.

Furthermore, we observe that factual questions obtain the highest scores, while synthesis questions obtain lower scores for each measure. We

model	type	R-1	R-L	BLEU
llama-FT	course	0.50	0.43	21.88
	definition	0.52	0.48	22.39
	factual	0.66	0.58	35.29
	synthesis	0.43	0.35	13.87
mistral-FT	course	0.37	0.31	11.88
	definition	0.44	0.39	14.85
	factual	0.55	0.49	23.08
	synthesis	0.33	0.26	7.94

Table 6: N-grams based metrics performances for the different type of question for the different models and configuration

can interpret this result in two ways: firstly, synthesis questions are much more difficult to answer correctly; secondly, synthesis answers leave more freedom in the words chosen, i.e. different formulations of the answer are possible. By contrast, answering factual questions is straightforward, as the range of vocabulary is limited. However, we cannot easily confirm or refute these hypotheses without human evaluation.

6.4. Human evaluation of the answers

Evaluate quality of the answer In text generation, it is not easy to consider automatic evaluation, even if we get reference answers, as language is the dress of thought, and we have many ways to express the same idea. Consequently, we designed metrics based on the human judgment. For the following experiments we ask six educated annotators to evaluate the answer according to binary criteria :

- **SYN**: Is the answer syntactically correct?
- **UND**: Is the answer semantically correct?
- **COR**: Is it the correct answer ?
- **CTX**: Does the answer use the document given or retrieved to produce the answer without adding any additional information?
- **PAR**: Does the answer miss some information or could be improved?

We first ask the annotators to judge 20 answers following the criteria below (given paragraphs context and the question) and evaluate an agreement. The results are reported in the table 7[7], if most of the annotators agreed on the different criteria (from medium to high agrement), the criterion "CTX" seems to be the exception. A hypothesis is that

[7]For Fleiss Kappa measure, a score between 0.2 and 0.4 is considered as a slight agreement, from 0.4 to 0.6 to medium agreement and above to high agreement

SYN	UND	COR	CTX	PAR
.80	.51	.63	.34	.52

Table 7: Fleiss kappa computed across five different criterion (binary criterion) on 20 annotations

Model	SYN	UND	COR	CTX	PAR
L-FT	87.5	94.2	65.0	74.2	47.5
L-FTR	89.2	89.2	49.2	63.3	47.5
M-FT	24.2	31.7	35.0	43.3	60
M-FTR	25.0	38.3	25.8	43.3	65.8

Table 8: Human evaluation scores obtained on Llama2-7b and Mistral-7b, with source context or BM25 document retrieval (in %).

annotators did not agree on what is used from the context to answer the question or consider the criterion irrelevant for a wrong answer.

We report in the table 8 the average of those criteria for 120 answers evaluated for each model (480 different questions were evaluated in total, with 80 different questions by annotator). The evaluation suggests that indeed LLAMA model is better for the task (at least in the French language) as it has a much higher score than Mistral. Particularly, different experiments could be designed in future work to improve Mistral answers, such as changing the prompt or using more French data for fine-tuning (in the pre-trained model we fine-tuned, no French instruction-based data were considered). Mistral model, in the current training setting and for the current task, is not relevant. Naturally, having correct documents not only leads to improved answers (65% vs 49.2%) but also eases the incorporation of context into answer generation (74.2 vs 63.3).

Evaluation by type of question An important question remains: how the type of question has an impact on the model's performance. In the table 8, we reported the average 'yes' answer from the annotators for each criterion according to the question type. As anticipated, utilizing BM25 prior to generation results in lower scores for the question correctness criterion, which is likely attributed to the presence of out-of-topic retrieved documents. Interestingly, table 9 reveals that the performances of the models remain consistent whatever the type of the question. It contrasts with the automatic evaluation where the performance decreases with the difficulty of the question. Especially, for synthesis questions, which obtain the same performance or better one than other types. As a result, we can assume that, for synthesis questions, numerous answers may exist that do not rely on specific word choices, while for other question types, it is more likely that the answers use a closer vocabulary. A limitation of our approaches is the incomplete or improvable answers according to the PAR criterion.

type	model	UND	COR	CTX	PAR
Course	L-FT	96.2	67.9	79.2	0.0
	L-FTR	92.5	54.7	75.5	20.8
Definition	L-FT	88.5	65.4	73.1	0.0
	L-FTR	88.5	57.7	57.7	26.9
Factual	L-FT	95.7	60.9	82.6	4.3
	L-FTR	91.3	39.1	60.9	21.7
Synthesis	L-FT	94.4	61.1	50.0	0.0
	L-FTR	77.8	33.3	38.9	33.3

Table 9: Human evaluation by question type(%).

Quality of the questions In addition to the previous annotation, we also asked the six annotators to provide information on the quality of the questions. During the annotation, the annotators had to label if the question was incorrect and not-understandable, incorrect but meaningful (small correction could make the question correct), or correct. 3.75% of the questions fall into the first category, those questions have to be cropped from the dataset. The second category contains 13.1% of the evaluated set, for this category we plan to organize a subsequent annotation campaign to rectify and improve those data (correction of questions and answers text). Consequently, 84% of the questions are today considered to be of high quality.

7. Discussion

In this work, we presented a new question-answering corpus for education in the French language. Our main focus is on developing educational content that closely resembles what a teacher can create, encompassing both questions and answers. In the different analyses of the corpus, we highlight the characteristics of the collected data, showing that the question and answer distribution is significantly different from mainstream question-answering dataset. Notably, a large part of the answers to the questions rely on explanations rather than straightforward factual answers.

The primary objective is to create a corpus suitable for designing systems that assist teachers to design courses, such as question and answer relying on educational material. To verify its adequacy, we developed a Retrieval Augmented Framework, using a retrieval and an answer generation system powered by large language models. We analyzed performances using both human and automatic evaluation, in order to offer insights and information regarding the corpus's quality and its suitability for training the different components of the RAG framework.

However, while the quality and the educational

focus of the corpus seem promising for further approaches, some issues remain. Indeed, the evaluation reveals that a few questions are incorrect. Although, the human evaluation showed that most of them can be easily reformulated. We intend to initiate a new annotation campaign to rectify those issues and ensure the highest quality questions.

Additionally, our exploration on other courses revealed that textual information is not enough to answer complex questions. Particularly in biology, most of the courses are based on schematics. We therefore intend to extend this corpus to design answers and questions that encompass both images and textual information.

8. Acknowledgement

This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011013721 made by GENCI. The data collection and annotation was financed by the SATT-Paris-Saclay and the company Stellia.

9. Bibliographical References

Elie Antoine, Jeremy Auguste, Frédéric Béchet, and Géraldine Damnati. 2022. Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, TALN-RECITAL 2022, Avignon, France, June 27 - July 1, 2022*, pages 104–115. ATALA.

Frederic Béchet, Cindy Aloui, Delphine Charlet, Geraldine Damnati, Johannes Heinecke, Alexis Nasr, and Frederic Herledan. 2019. CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*, pages 185–194. ATALA.

Frédéric Béchet, Ludivine Robert, Lina Rojas-Barahona, and Géraldine Damnati. 2022. CalorDial : a corpus for Conversational Question Answering on French encyclopedic documents. In *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe)*, volume 3178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. Fquad: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1193–1208. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine J.-P. Tixier, and Michalis Vazirgiannis. 2021. Barthez: a skilled pretrained french sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9369–9390. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han.

2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1051–1068. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: building a native french question-answering dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5481–5490. European Language Resources Association.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2479–2490. European Language Resources Association.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7203–7219. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.