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

Thomas Gerald*, Louis Tamames*, Sofiane Ettayeb, Patrick Paroubek and Anne Vilnat

May 3, 2024

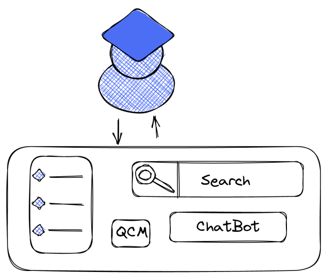
Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS

Introduction

Motivations:

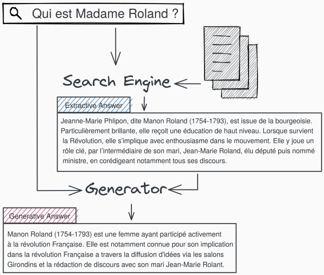
Develop a teaching assistant (collaboration with the company Stellia)

- Recommend class materials (according to student level)
- Propose tools to help student to retains knowledge and to progress
  - Multiple Choice Questions
  - Course questions and their answers
- Answer student's questions (when teacher is not available or to help comprehension)



Introducing CQuAE: A corpus for Educational purpose

Introducing CQuAE: Final objectives and issues



Jeanne-Marie Phlipon, dite Manon Roland (1754-1793), est issue de la bourgeoisie. Particulièrement brillante, elle reçoit une éducation de haut niveau. Lorsque survient la Révolution, elle s’implique avec enthousiasme dans le mouvement. Elle y joue un rôle clé, par l’intermédiaire de son mari, Jean-Marie Roland, élu député puis nommé ministre, en corédigeant notamment tous ses discours.

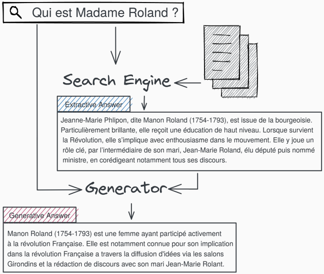
Manon Roland (1754-1793) est une femme ayant participé activement à la révolution Française. Elle est notamment connue pour son implication dans la révolution Française a travers la diffusion d'idées via les salons Girondins et la rédaction de discours avec son mari Jean-Marie Rolant.

Our question answering system:

Set up a question answering system relying on:

- Answering questions for helping students
- Generate answers that are grounded relying, on a restricted set of educational materials
- Check/verify the information, ensure the system will not create counter-factual answers (qualities)

Introducing CQuAE: Final objectives and issues



Jeanne-Marie Phlipon, dite Manon Roland (1754-1793), est issue de la bourgeoisie. Particulièrement brillante, elle reçoit une éducation de haut niveau. Lorsque survient la Révolution, elle s'implique avec enthousiasme dans le mouvement. Elle y joue un rôle clé, par l'intermédiaire de son mari, Jean-Marie Roland, élu député puis nommé ministre, en corédigeant notamment tous ses discours.

Manon Roland (1754-1793) est une femme ayant participé activement à la révolution Française. Elle est notamment connue pour son implication dans la révolution Française a travers la diffusion d'idées via les salons Girondins et la rédaction de discours avec son mari Jean-Marie Rolant.

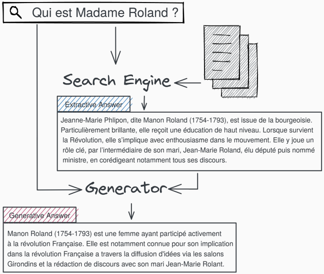
Our question answering system:

Set up a question answering system relying on:

- Answering questions for helping students
- Generate answers that are grounded relying, on a restricted set of educational materials
- Check/verify the information, ensure the system will not create counter-factual answers (qualities)

→ **Need a training corpus**

Introducing CQuAE: Final objectives and issues



Jeanne-Marie Phlipon, dite Manon Roland (1754-1793), est issue de la bourgeoisie. Particulièrement brillante, elle reçoit une éducation de haut niveau. Lorsque survient la Révolution, elle s'implique avec enthousiasme dans le mouvement. Elle y joue un rôle clé, par l'intermédiaire de son mari, Jean-Marie Roland, élu député puis nommé ministre, en corédigeant notamment tous ses discours.

Manon Roland (1754-1793) est une femme ayant participé activement à la révolution Française. Elle est notamment connue pour son implication dans la révolution Française a travers la diffusion d'idées via les salons Girondins et la rédaction de discours avec son mari Jean-Marie Rolant.

Our question answering system:

Set up a question answering system relying on:

- Answering questions for helping students
- Generate answers that are grounded relying, on a restricted set of educational materials
- Check/verify the information, ensure the system will not create counter-factual answers (qualities)

→ **Need a training corpus**

Available corpus:

- **FQuAD, SQuADFR, Piaf, SQuAD** [MMWT20, KLB+20, RZLL16] mainly factual and simple answers (e.g. "what country is Normandy located?")
- **Natural Questions, HotpotQA, ...** [KPR+19, YQZ+18] complex multi-hop questions with short answers (HotpotQa), answers not only focusing on the question

Introducing CQuAE: A corpus Issue

Exisiting problems in available corpora:

- **To simple questions**: Most of the responses rely on a very short answer
- **No diversity in the questions**: Most of the answers rely on capturing entity (name, place, date)
- **Not relying on education**: Document are not based upon education material
- **Most of the corpora are in English**

Introducing CQuAE: A corpus Issue

Exisiting problems in available corpora:

- **To simple questions**: Most of the responses rely on a very short answer
- **No diversity in the questions**: Most of the answers rely on capturing entity (name, place, date)
- **Not relying on education**: Document are not based upon education material
- **Most of the corpora are in English**

This work

→ **Propose a new corpus to tackle previous corpora problems**

Introducing CQuAE: Build a new corpus

Contributions:

Introducing CQuAE: Build a new corpus

Contributions:

- **Propose a new french question-answering corpus on educational content for teaching assistant purpose (history, geography, biology)**

Introducing CQuAE: Build a new corpus

Contributions:

- **Propose a new french question-answering corpus on educational content for teaching assistant purpose (history, geography, biology)**
- **Complexes questions oriented annotation (prefer explanation as answer than entity or one fact)**

Introducing CQuAE: Build a new corpus

Contributions:

- **Propose a new french question-answering corpus on educational content for teaching assistant purpose (history, geography, biology)**
- **Complexes questions oriented annotation (prefer explanation as answer than entity or one fact)**
- **Study the corpus characteristics and a comparison with standard datasets**

Introducing CQuAE: Build a new corpus

Contributions:

- **Propose a new french question-answering corpus on educational content for teaching assistant purpose (history, geography, biology)**
- **Complexes questions oriented annotation (prefer explanation as answer than entity or one fact)**
- **Study the corpus characteristics and a comparison with standard datasets**
- **Evaluate the corpus in a Retrieval Augmented Generation framework**

Introducing CQuAE: Resources

Repères

L'engrenage des alliances

Compétence Je me repère dans le temps et dans l'espace

À la fin du XIXe siècle, le nationalisme des pays européens aboutit à la mise en place de deux systèmes d'alliances. La Triple-Entente (France, Royaume-Uni, Russie) s'oppose à la Triple-Alliance (empires allemands et austro-hongrois, Italie). Après l'assassinat, le 28 juin 1914 à Sarajevo, de l'archiduc François-Ferdinand, héritier de l'empire d'Autriche-Hongrie, le conflit éclate et s'élargit progressivement.

1 La Première Guerre mondiale (1914-1918)

Notions

Une guerre totale

Mobilisation humaine, économique, culturelle et morale de toute la société pour gagner la guerre.

1. Doc. 3 et affiche ci-contre Relevez sur ces deux affiches plusieurs formes d'implication des populations dans la guerre.
2. Doc. 3 Que représente la bête ? Qui tient-elle entre ses bras ? Que tient-elle dans son autre main ?
3. Doc. 3 et affiche ci-contre Quel est le but de ces affiches ?

Affiche de propagande du gouvernement français pendant la Première Guerre mondiale.

2 Une guerre mondiale

Des soldats algériens dans l'Oise, 1917.

3 La propagande

« Détruisez cette bête furieuse, engagez-vous ! », affiche américaine, 1917.

Vocabulaire

Le front : zone dans laquelle les combats ont lieu.
Une guerre mondiale : conflit qui implique des États de plusieurs continents. En 1914, la guerre débute en Europe, puis entraîne les États-Unis et le Moyen-Orient. Des soldats coloniaux (Afrique, Asie) sont mobilisés dès le début du conflit. Des batailles ont également lieu dans l'Atlantique.
Le nationalisme : idéologie politique qui repose sur l'affirmation de la supériorité de la nation.

CHAPITRE 1 • La Première Guerre mondiale LLS.fr/HG3P33 33

Collecting the resources:

- **lelivrescolaire**[a]**:** Schoolbook (High and middle school) under creative common license
- **Wikipedia:**
  - Retrieve pages related to lelivrescolaire title
  - Split the document (section)

Present to annotators each document or section to annotate.

[a] lelivrescolaire.fr

Topics and grades:

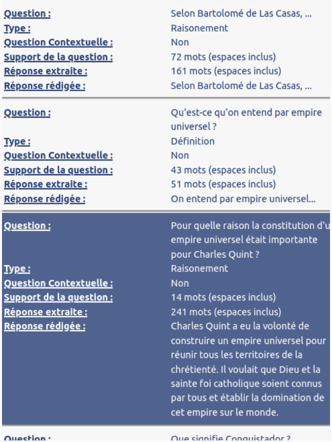
According to french education program (equivalence with US given)

- History - 11th, 10th, 7th and 6th grades (première, seconde, cinquième, sixième)
- Geography - 11th, 10th, 7th and 6th grades
- EMC - 7th and 6th grades
- Biology/geology - 10th and 7th grades

[0] Image from https://www.calameo.com/read/000596729e1afc25319d5

Introducing CQuAE: Annotation

1. **The question**: a question formulated by the annotator regarding the document;
2. **The type of question**: we propose four classes: factual, definition, course, and synthesis;
3. **The question support**: a passage in the text serving as support for generating a question (a short answer);
4. **The answer elements**: a selection of several passages allowing to answer the different elements of the question;
5. **The written answer**: an answer written by the annotator summarizing the different elements of the answer.



Question :	Selon Bartolomé de Las Casas, ...
Type :	Raisonement
Question Contextuelle :	Non
Support de la question :	72 mots (espaces inclus)
Réponse extraite :	161 mots (espaces inclus)
Réponse rédigée :	Selon Bartolomé de Las Casas, ...
Question :	Qu'est-ce qu'on entend par empire universel ?
Type :	Définition
Question Contextuelle :	Non
Support de la question :	43 mots (espaces inclus)
Réponse extraite :	51 mots (espaces inclus)
Réponse rédigée :	On entend par empire universel...
Question :	Pour quelle raison la constitution d'u empire universel était importante pour Charles Quint ?
Type :	Raisonement
Question Contextuelle :	Non
Support de la question :	14 mots (espaces inclus)
Réponse extraite :	241 mots (espaces inclus)
Réponse rédigée :	Charles Quint a eu la volonté de construire un empire universel pour réunir tous les territoires de la chrétienté. Il voulait que Dieu et la sainte foi catholique soient connus par tous et établir la domination de cet empire sur le monde.
Question :	Que signifie Conquistador ?

Introducing CQuAE: The question type

Having questions requiring different levels of expertise, depending on the type of information needed to answer them

Introducing CQuAE: The question type

Having questions requiring different levels of expertise, depending on the type of information needed to answer them

- **Factual**: The answer is a fact or a list of facts (event, person, location, date...).

Introducing CQuAE: The question type

Having questions requiring different levels of expertise, depending on the type of information needed to answer them

- **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.

Introducing CQuAE: The question type

Having questions requiring different levels of expertise, depending on the type of information needed to answer them

- **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.

Introducing CQuAE: The question type

Having questions requiring different levels of expertise, depending on the type of information needed to answer them

- **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.

Introducing CQuAE: The question type

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...

Introducing CQuAE: Annotation campaign

Two groups

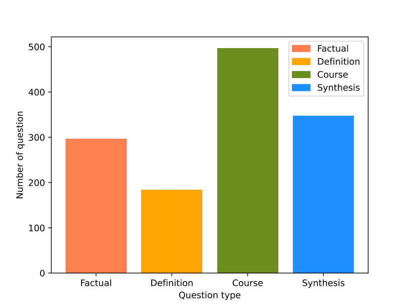
- **Group A:** No specific teaching backgrounds but educated
- **Group B:** Knowledgeable and a specific educational background

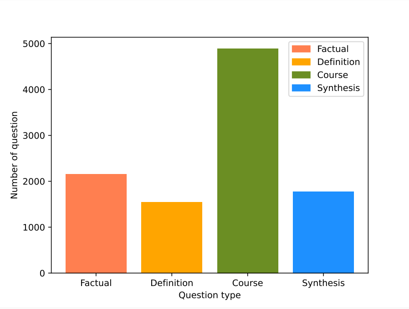
Introducing CQuAE: Annotation campaign

Two groups

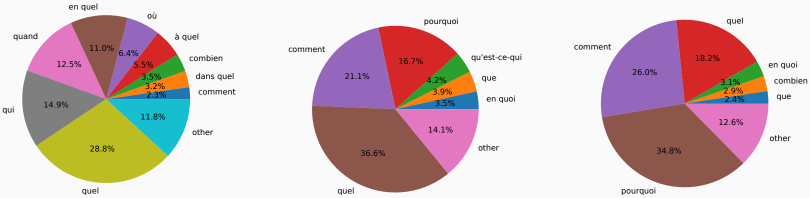
- **Group A:** No specific teaching backgrounds but educated
- **Group B:** Knowledgeable and a specific educational background

Qu. Type	Group A	Group B	Total
Course	4 784	490	5 274
Factual	2 106	294	2 400
Synthesis	1 756	338	2 094
Definition	1 506	181	1 687
Total	10 152	1 303	11455



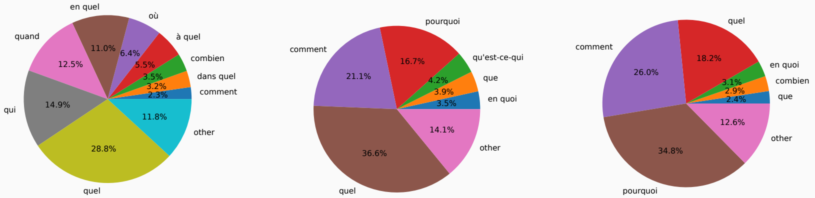


Introducing CQuAE: Analysis of the question types



(a) Factual **(b)** Course **(c)** Synthesis

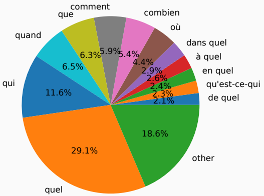
Introducing CQuAE: Analysis of the question types



(a) Factual

(b) Course

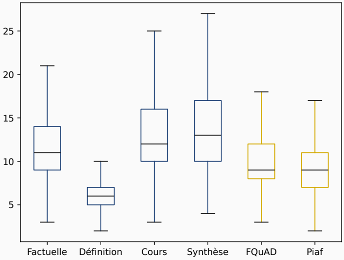
(c) Synthesis



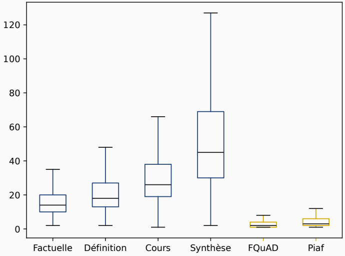
(a) FQuAD question word distribution

- FQuAD question word distribution close to Factual in our corpus
- Disimilar distribution (leads to more complex answer) for course and synthesis

Introducing CQuAE: Comparing question/answer length



(a) Question length



(b) Answer length

Question and Answer length

- Sligthly Longer questions than FQuAD and PIAF (except for defintion)
- **Largely Longer answers** → redacted answers and rarely only one entity (place, person, date, . . .)

Introducing CQuAE: Difficult questions (cherry picked)

Difficult Questions: QAE-A

- What does the expression "power is using a glass chisel to sculpt marble" mean? What does Louise Michel think of this expression? **Que veut dire l’expression "le pouvoir, c’est se servir d’un ciseau de verre pour sculpter le marbre" ? Que pense Louise Michel de cette expression ?**
- "Your reign passes like that of the Tyrants." What does Olympe de Gouges mean? How does her defense go beyond herself? **"Votre règne passe comme celui des Tyrans". Que veut dire Olympe de Gouges ? En quoi sa défense dépasse sa seule personne ?**

Difficult Questions: QAE-B

- According to the excerpt from "Germany since the war of 1866" by Emile de Laveleye, how would you summarize his viewpoint on the unity of Austria? **D’après l’extrait de "L’allemagne depuis la guerre de 1866" d’Emile de Laveleye, comment résumeriez-vous son point de vue sur l’unité de l’Autriche ?**
- How do institutions guarantee and protect freedoms in France and Europe? **Comment les institutions garantissent et protègent-elles les libertés en France et en Europe ?**

First experiments using RAG framework

CQuAE: Experiments

Evaluate the corpus to generate grounded answer:

Use of the Retrieval Augmented Framework (search and summarize)

- How can we retrieve relevant documents (or paragaphs) from collected questions?
- How State-of-the-art models perform on this dataset?
- Is difficulty dependent to the question type?

	Train	Validation	Test
N-que	10 490	407	558
Fact	21%	22%	19%
Def	14%	14%	19%
course	46%	45%	45%
synthesis	18%	19%	17%
SB	46%	54%	45%

CQuAE: Retrieval approach

Retrieval Approach:

Retrieve document(s) (paragraphs) on which the question was created

- **BM25:** TF-IDF based approach
- **DPR:** Using LLM to encode both query and documents (Dense representation)

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

Table 1: Ranking performances on our corpus for the different approaches

- Better performances from the "naive" approach (BM25)
- Half of the target documents are not retrieved first (*P*@1)
  → But retrieved documents still can be relevant

CQuAE: Generative approach

Generate answers

- From a question and documents → generate an answer to the question
- Three configurations: zero-shot (no fine-tuning), fine-tuned (on our training set), and, retrieval (using document retrieved from BM25)
- Two models: LLAMA2-7b [TLI+23] and Mistral-7b [JSM+23]

Conf	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

- Mistral better in zero-shot (no fine-tuning), LLAMA2 better elsewhere
- Lower score with retrieval but low difference → documents not belonging to target mostly relevants?
- **Rouge and bleu not very informative**

[0] We noticed that changes in prompt and generation parameters can change drastically the performances

CQuAE: Human evaluation criterion

Automatic evaluation metrics are not sufficient:

- Low BLEU/ROUGE score does not mean that answer is incorrect
- Does the response contain the answer to the question without missing or additional information?

CQuAE: Human evaluation criterion

Automatic evaluation metrics are not sufficient:

- Low BLEU/ROUGE score does not mean that answer is incorrect
- Does the response contain the answer to the question without missing or additional information?

Human evaluation is necessary (binary criterion):

- **UND**: Is the answer semantically correct?
- **COR**: Is it a 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?

CQuAE: Evaluation and question type

Evaluation campaign:

6 educated evaluators and 120 answers evaluated for each model

CQuAE: Evaluation and question type

Evaluation campaign:

6 educated evaluators and 120 answers evaluated for each model

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

Table 2: Human evaluation by question type(%).

CQuAE: Evaluation and question type

Evaluation campaign:

6 educated evaluators and 120 answers evaluated for each model

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

Table 2: Human evaluation by question type(%).

- FTR (with document retrievied by BM25) get lower performances

CQuAE: Evaluation and question type

Evaluation campaign:

6 educated evaluators and 120 answers evaluated for each model

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

Table 2: Human evaluation by question type(%).

- FTR (with document retrievied by BM25) get lower performances

  → The LLM uses the source to build the answer (it is less clear for the definition question type)

CQuAE: Evaluation and question type

Evaluation campaign:

6 educated evaluators and 120 answers evaluated for each model

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

Table 2: Human evaluation by question type(%).

- FTR (with document retrievied by BM25) get lower performances
  - → The LLM uses the source to build the answer (it is less clear for the definition question type)
  - → Some of the retrieved documents are irrelevant

CQuAE: Conclusion

- Create a new corpus for question-answering for educational purpose
- Compare CQuAE corpus to factual question-answering dataset
- Different experiments validating the relevance of the corpus

Remaining issues:

Some of the questions are irrelevant or contain errors:
→ 3.75% are irrelevant (estimated)
→ 13% are relevant but contain errors (estimated)

- In which city did Prussia lose to Austria?
  Dans quelle ville la Prussie a-t-elle perdu contre l'Autriche?
- To whom is Austria not attached?
  À qui n'est pas attaché l'Autriche?
- Against whom did Austria lose in 1866?
  Contre qui l'Autriche a-t-elle perdu en 1866?

[0] Data and scripts available at `https://gitlab.lisn.upsaclay.fr/gerald/cquae`

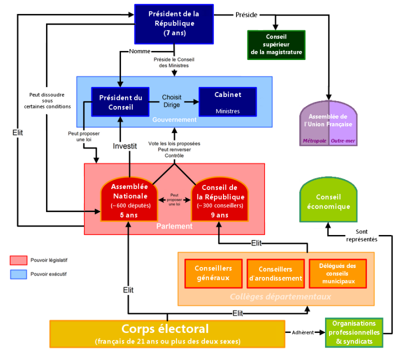
CQuAE: Next step(s)

- **Improvement on the quality of the corpus:** A second version of the corpus with revisions on questions and answers

[0] Image from https://fr.wikipedia.org/wiki/Quatri%C3%A8me_R%C3%A9publique_(France)

CQuAE: Next step(s)

- **Improvement on the quality of the corpus:** A second version of the corpus with revisions on questions and answers
- **Taking into account different modalities:** Schoolbooks often rely on image graphics
  - Extract information from schematics
  - Answer questions from text and schematics



Président de la République (7 ans)
Préside
Conseil supérieur de la magistrature
Nomme
Préside le Conseil des Ministres
Peut dissoudre sous certaines conditions
Président du Conseil
Choisit Dirige
Cabinet
Ministres
Gouvernement
Assemblée de l'Union Française
Métropole
Outre-mer
Elit
Peut proposer une loi
Investit
Vote les lois proposées Peut renverser Contrôle
Assemblée Nationale (~600 députés) 5 ans
Peut proposer une loi
Conseil de la République (~300 conseillers) 9 ans
Parlement
Conseil économique
Sont représentés
Elit
Pouvoir législatif
Pouvoir exécutif
Conseillers généraux
Conseillers d'arondissement
Délégués des conseils municipaux
Collèges départementaux
Elit
Elit
Corps électoral (français de 21 ans ou plus des deux sexes)
Adhèrent
Organisations professionnelles & syndicats

[0] Image from https://fr.wikipedia.org/wiki/Quatri%C3%A8me_R%C3%A9publique_(France)

Thank you

References i

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, *Mistral 7b*, CoRR **abs/2310.06825** (2023).

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, FrÃ©dÃ©ric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano, *Project piaf: Building a native french question-answering dataset*, Proceedings of The 12th Language Resources and Evaluation Conference, May 2020.

References ii

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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, *Natural questions: A benchmark for question answering research*, Transactions of the Association for Computational Linguistics (2019).

d'Hoffschmidt Martin, Vidal Maxime, Belblidia Wacim, and Brendlé Tom, *FQuAD: French Question Answering Dataset*, arXiv e-prints (2020).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, *Squad: 100, 000+ questions for machine comprehension of text*, EMNLP, The Association for Computational Linguistics, 2016.

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, *Llama: Open and efficient foundation language models*, CoRR **abs/2302.13971** (2023).

References iii

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning, *Hotpotqa: A dataset for diverse, explainable multi-hop question answering*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018 (Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, eds.), Association for Computational Linguistics, 2018, pp. 2369–2380.