On the Feasibility of LLM-based Automated Generation and Filtering of Competency Questions for Ontologies

Zola Mahlaza[1], C. Maria Keet[1,2], Nanée Chahinian[3], Batoul Haydar[3]
[1]Department of Computer Science, University of Cape Town, South Africa,
zmahlaza@cs.uct.ac.za
[2]Meaningfy SARL, Lintgen, Luxembourg
[3]HSM, Univ Montpellier, IRD, CNRS, Montpellier, France

Abstract

Competency questions for ontologies are used in a number of ontology development tasks. The questions' sentences structure have been analysed to inform ontology authoring and validation. One of the problems to make this a seamless process is the hurdle of writing good CQs manually or offering automated assistance in writing CQs. In this paper, we propose an enhanced and automated pipeline where one can trace meticulously through each step, using a mini-corpus, T5, and the SQuAD dataset to generate questions, and the CLaRO controlled language, semantic similarity, and other steps for filtering. This was evaluated with two corpora of different genre in the same broad domain and evaluated with domain experts. The final output questions across the experiments were around 25% for scope and relevance and 45% of unproblematic quality. Technically, it provided ample insight into trade-offs in generation and filtering, where relaxing filtering increased sentence structure diversity but also led to more spurious sentences that required additional processing.

1 Introduction

The use of Competency Questions (CQs) for ontology scoping, development, and validation is well-established since its introduction in 1996 (Uschold and Gruninger, 1996), as illustrated in, e.g., (Alharbi et al., 2023; Bezerra and Freitas, 2017; Bezerra et al., 2013; Keet and Lawrynowicz, 2016; Suarez-Figueroa et al., 2008; Thiéblin et al., 2018). Authoring CQs is not trivial and a question's wording may be problematic for a number of reasons (Khan and Keet, 2024). Therefore, effort has gone into CQ authoring assistance. Early efforts went into creating a Controlled Natural language to assist writing, called CLaRO (Keet et al., 2019; Antia and Keet, 2021), but with the advances and popularization of Large Language Models (LLMs), the allure of LLM-assisted authoring has gained traction (Alharbi et al., 2024b). Variants include retrofitting CQs onto an existing ontology using, e.g., a prompting-based approach (Alharbi et al., 2024a), or generating CQs for a prospective ontology yet to be developed, which can be done with training or fine-tuning (Antia and Keet, 2021) or prompting (Pan et al., 2025).

While retroactively generating CQs for an existing ontology has usage scenarios relevant for the ontology development lifecycle, we are interested in the scenario where the ontology is yet to be developed, irrespective of, though possibly including, ontology reuse, such as for scoping the subject domain and therewith formulating the requirements. Structured CQs can then feed into semi-automated ontology authoring (Wisniewski et al., 2021) and querying (Keet and Lawrynowicz, 2016; Wisniewski et al., 2019). The broad question it raises is *how to automate and obtain relevant CQs and to do this in such a manner that the CQs can be traced to the source?*. AgOCQs (Antia and Keet, 2021) aims to cater for this scenario, using the T5 LLM, the SQuAD dataset for fine-tuning, filtering with the CLaRO v2 CNL for CQs, and a semantic filtering step. However, it was evaluated with only one use case, a very small corpus of 7 scientific articles, and the effects of the different steps in the pipeline are unclear as only the final output was evaluated. Our aims are to focus on fully automating all aspects of that pipeline from text extraction, generation, and filtering, in a traceable manner, possibly enhance it where promising, and test it on another subject domain. Specific questions we seek to answer are:

1. Is the AgOCQs pipeline effective for use cases in other domains than it was tested on (COVID-19)?
2. Is AgOCQs effective on other types documents, i.e., not just scientific articles, but also standards and guidelines?

3. What is the effect of different corpus size on the number and quality of the CQs generated?

4. What exactly is the contribution of each filtering step on AgOCs's output?

5. What is the effect of the SQuAD training set on the quality of the output?

To answer this, we refactored the Jupyter notebook from (Antia and Keet, 2023) and ran preliminary tests to answer RQ-3. In the first experiment, we ran the pipeline with two mini-corpora, one consisting of guidelines and another with scientific documents, and evaluated the generated questions with two domain experts and an ontologist, to answer RQ-2 and RQ-1, and aimed to answer RQ-4 and RQ-5. In Experiment 2 we modified the pipeline in a number of ways to obtain more fine-grained insights and answers to RQ-4, RQ-5, and RQ-1.

The questions outputted by the pipeline for both experiments were around 25% for scope and relevance out of the total evaluated, and when within scope, then they were for 69-75% relevant, with quality from an ontological viewpoint varying between 53% and 40% as acceptable or good CQ for ontologies. This was obtained with full automation, cf. the original AgOCQs that required manual curation. The tracing in the automation provided ample insight into trade-offs. Important steps affecting the process are the SQuAD training data set and the filtering step with the CLaRO CNL, and various minor gains were obtain with grammar checking, English checking, and an additional conceptual filter that removed CQs appropriate for conceptual data models and the ABox rather than ontologies.

In the remainder of the paper, we describe the materials and methods in Section 2, present the results in Section 3, and discuss and conclude in Sections 4 and 5.

2 Methodology

For purposes of being self-contained, this section will first summarise AgOCQs, and subsequently the materials and methods for the two experiments.

2.1 Background: AgOCQs

The first step in AgOCQs is extracting the domain text corpus and to preprocess it with Spacy for sentence extraction and stop word removal (Honnibal and Montani, 2017) and regualr expressions to produce cleaned data. This is fed to the T5 base model (Raffel et al., 2020) that is pre-trained with the SQuAD dataset (context and question) as source task. It outputs the context texts and questions, which is "de-cluttered" with the Sentence Transformer model (Reimers and Gurevych, 2019), which includes removing duplicates.

The output is analysed on sentence structure using Wisniewski's code (Potoniec et al., 2020; Wisniewski et al., 2019), resulting in patterns of text with entity and predicate chunks, which are then compared against the patterns that were at the basis of CLaRO v2 (Antia and Keet, 2021; Keet et al., 2019). If they match fully, the generated question is a candidate CQ.

2.2 Preparation

The first step consisted of analysing the CQ generation pipeline of AGoCQs, both on what was reported in (Antia and Keet, 2023) and the associated Jupyter notebook, with preparations and pre-experimentation. This involved automating all aspects to further reduce the manual curation found in the pipeline and increasing the maximum number of training epochs to 2.

The updated pipeline automatically extracts text from each PDF file using PyPDF2[1] and ach file is then tokenized to obtain sentences using Spacy[2] (Honnibal and Montani, 2017). The pipeline then generates three questions for each sentence. Each question is cleaned up in a simple manner (e.g., removing the text generating model's prefix and ensuring that each output ends with a question mark), abstracts the questions using the source code from (Wisniewski et al., 2019) to obtain abstract patterns of the questions, filters out questions whose abstract patterns are not found in CLaRO v2 (Antia and Keet, 2021), and eliminates questions that are too similar to each other. A question is too similar to another if there exists another question whose cosine similarity exceeds 0.75, as determined using representations obtained using the all-MiniLM-L6-v2[3] model from the Sentence Transformer (Reimers and Gurevych, 2019) library.

Traceability was also added so that during running the pipeline, it can generate a csv file after completing each step. This enables tracing forward and backward, i.e., from a paragraph in the mini-corpus to question generated, its chunking, its matching (or not) with a CLaRO v2 template,

[1] https://pypdf2.readthedocs.io/en/3.x/
[2] https://spacy.io/
[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

and its keeping or removing from the set thanks to the semantic similarity check. This also enables answering the question about what a good size of the mini-corpus may be.

We dub this enhanced version of Antia and Keet's algorithm AgOCQs+, which is depicted graphically in the top-half of Figure 1.

2.3 Experiment 1: AgOCQs+ With Another Subject Domain

The aim of the first experiment is twofold: execute it on a different subject domain to test generalizability of the approach and gain insights into the effects of each step in the pipeline to serve as potential sources of improvement.

As subject domain, we choose wastewater and stormwater networks, because the ontology is under development by collaborators on a project (Keet et al., 2025) and such physical network infrastructure is distinct from knowledge about COVID-19. In addition, its aim was to 'ontologise' sewer network standards and guidelines, which is a starkly different setting from rapidly evolving knowledge about a new pathogen and symptoms it generates. One domain expert collected standards and wastewater guidelines that were in English and freely available online from the EU (Ireland), the Americas (Canada), and Africa (South Africa), totalling 4 documents, and distinct from the guidelines already used in the ontology development (described below). The same domain expert also selected 4 scientific articles in the subject domain of the ontology under development, to examine the possibility of mini-corpus genre effects on AgOCQs+.

Regarding examining the effects of each step, it is hoped we gain insight into aspects such as whether a question is justly discarded for indeed being the same or too similar, and how many, and any false positives or negatives due to CLaRO filtering.

Overall assessment also includes a domain expert evaluation. Its aim is to assess whether sufficient in-scope CQs are generated that are relevant for the ontology and that would be formalisable/answerable in an (at most) OWL 2 DL ontology. The main hypotheses were formulated as follows:

H1 Questions generated from the other (i.e., not yet used and in English) standards and guidelines will significantly more often be relevant than those generated from the scientific texts.

H2 Questions generated from the scientific texts will significantly more often be relevant than those generated from the guidelines.

H3 Scope and relevance percentages are in the same range as observed for the experiment with the COVID-19 CQ generation, and there will be more useful questions than useless ones.

H1 is motivated by the fact that the original plan was to ontologise the standards such that the ontology would be relevant also beyond RAEPA and INSPIRE, the geostandards used to build the ontology. H2, a converse of H1, may be argued for because standards have a myriad of text that is 'off-topic' for the ontology, which in-domain scientific papers are expected not to have. That is: there are different reasons why a mini-corpus in one or the other genre may, or may not, be effective. H3 is included because AgOCQs and AgOCQs+ are assumed to perform well regardless the subject domain.

The procedure for the human evaluation is as follows.

1. Select 200 candidate CQs from those generated from the standards and the scientific papers (100 from each set), and ensure the origin is not viewable by the participants in the excel sheet where the will enter the judgements.

2. Two domain experts annotate each question on it being within the scope of the domain of wastewater and stormwater (yes/no), and if yes, select for relevance for the SewerNet ontology (yes/partial/no), where 'partial' means that the question can become 'yes' after a slight tweak, i.e., the CQs are found relevant if SewerNet can answer them or can be extended to address them. For instance, questions about drinking water and documents are out of scope, a question about wastewater quality measurement is within scope but not relevant, and questions about a combined wastewater pipe's diameter or a manhole cover are both within scope and relevant.

3. For all coded 'yes' on scope and relevance, one ontologist annotates whether the question is problematic as CQ or not, and if problematic, why. Problematic may be grammar, vague or imprecise terms, or concerning content inappropriate for (the TBox of) an ontology.

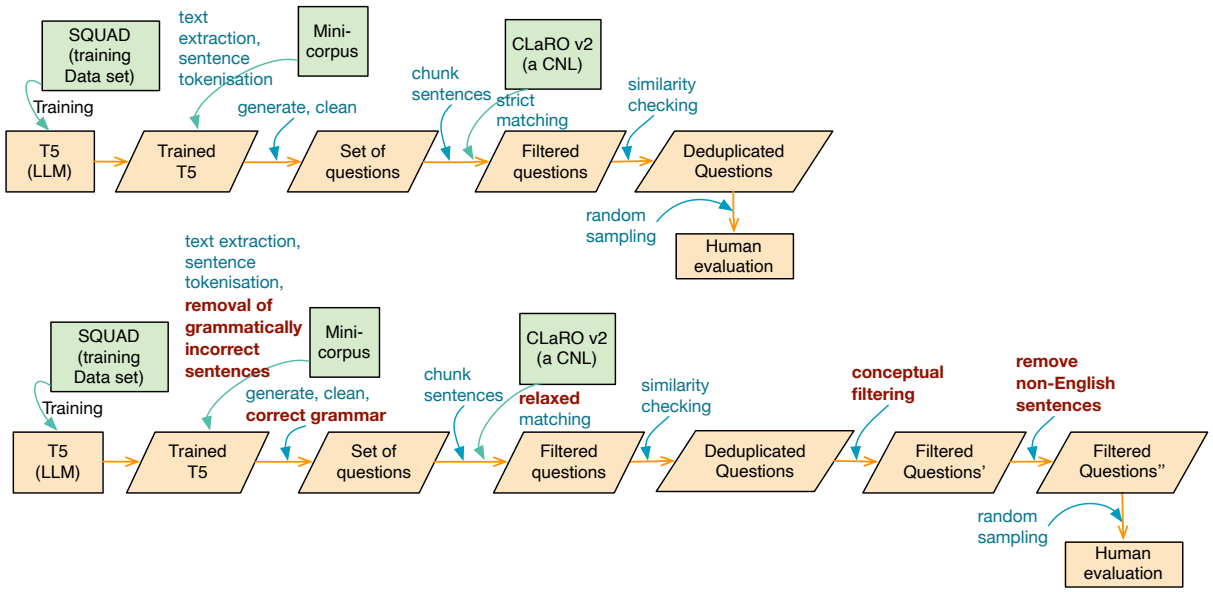

Figure 1: Automated AgOCQs+ pipeline (top) and AgOCQs++ pipeline (bottom) where the changes cf. AgOCQs+ are highlighted in bold maroon.

4. Compute descriptive statistics and inter-annotator agreement for the domain experts (if it is low, they will be asked to discuss their judgements). Agreement scores are determined using Cohen's Kappa (Cohen, 1960).

The outcomes of this experiment will inform a subsequent experiment.

The materials used are those listed in the preparation (Section 2.2) concerning the computational component. The two mini-corpora are available in the supplementary material and listed in Appendix A. The ontology considered for the relevance question is SewerNet[4] which describes the structure of sewer networks and their elements and qualities (Keet et al., 2025). The ontology is aligned with the DOLCE-lite foundational ontology and imports a few axioms from the Time ontology. The first corpus contains sewer network design guidelines from English-speaking countries (Canada, Ireland, and South Africa) and the second corpus contains articles published in water science journals with a Q1 SJR Rank, that were comparatively recent, which focused on the network itself either for modelling or asset management and IoT (see Appendix), whereas articles on ontology development or use for the domain were voluntarily discarded. For the human evaluation data and collection, MS Excel was used, and for analysis we computed the percentages of positively judged questions and measured agreement using Cohen's Kappa (Cohen, 1960).

2.4 Experiment 2: Permutations and extensions to AgOCQs+

The aims of the second round of experiments are to improve the quality of the output, to some extent on relevance for the domain experts, but more so that the quality of the questions should be good as CQs for ontologies, since a domain expert rating a question as relevant but it not being answerable by an (at most OWL 2 DL-formalised) ontology is of little use in automation of ontology requirements gathering and testing.

Assessing the intermediate outputs and the human evaluation, we devised a number of permutations and ran the adjusted algorithms again. They were the following variants:

P1 Exploring modifying SQuAD by filtering its questions by those matching the abstract form of the CLaRO v2 patterns.

E0 The Spacy sentence tokenizer returned 'sentences' of poor quality, at times. For instance, one of the sentences it returned was "H Sun P Thiagarajan R Anderson Owin OAuth Authorization Server" and its associated question (i.e., "What is the purpose of this website?") was judged as being out of scope by both experts and problematic since it is vague with respect to what it refers. Since such inputs can

[4] http://sewernet.msem.univ-montp2.fr/

lead to questions of poor quality, for this experiment we included a filter to remove sentences of poor grammatical quality. There is no readily available grammaticality metric that can be repurposed for such a task, to the best of our knowledge; thus we created a proxy metric. Specifically, we used the coedit-large model (Raheja et al., 2023) to correct each sentence and computed the Levenshtein distance[5] between the original and corrected version. We then normalized the scores via min-max, and filtered out values below 0.8 (initial) or 0.7 (final).

E1 Some of the generated questions were grammatically incorrect. For instance, there are missing braces in the question "What is the purpose of the Internet of Things IoT?". We added a grammar correction step to the raw generated questions using the coedit-large model (Raheja et al., 2023).

P2 The strict filter that removes questions that do not match the patterns on which CLaRO v2 is based resulted in a lack of diversity in the syntactic structure of questions. We modified the CLaRO filtering step to allow less than 100% matches by computing the edit distance of each abstracted sentence of a generated question, assigning the question the max edit distance, normalising the scores via min-max, and filtering out all questions whose scores are above 0.8 (initial) or 0.6 (final).

E2 We identified common issues that make questions invalid as competency questions for ontologies (e.g., ones that ask for instances rather than type-level knowledge). We removed all questions that include the following phrases: 'examples of' and 'name of'.

E3 The model sometimes generated questions that included German phrases (e.g., "Um, the Angebot von Wurmwaren in South Carolina?") or were repeating the same word (e.g., "A Arundel Arundel Arundel Arundel Arundel ..."). We added a module to remove texts whose probability of being English is less than 0.5 using Facebook's fasttext-language-identification[6] model.

E4 Add all of P2, P3, E1, and E2 to AgOCQs+, dubbed AgOCQs++.

Human evaluation was carried out on a random selection of 48 questions (6 per document) outputted from E4, as described for Experiment 1.

Additional materials used are those listed in P1-E4.

3 Results

The results are presented in order of the experiments.

3.1 Preparation Phase

The preparation stage resolved initial questions. First, it was deemed difficult to determine from the Jupyter notebook whether AgOCQs as reported in (Antia and Keet, 2023) effectively ran similarity filtering before or after the CLaRO filtering. Similarity now certainly happens after the CLaRO filtering. Second, questions had been raised about the size of the mini-corpus, and specifically whether 7 scientific papers would be sufficient. The new traceability features enabled this assessment, therewith answering RQ3, as follows.

Of the four standards and guidelines, some were processed to remove the cover page, glossary etc., and with a limit set to 100 sentences per document, the pipeline would have been generating questions from different parts of the documents. In order of file processing, and thus, eventually, discarding duplicates, it largely exhausted generating distinct and new candidate CQs after processing three files. Specifically, 71 CQs were traced back to the 'Wastewater Code of Practice', 55 were generated from to the 'Wastewater' document, and 69 additional questions from the 'Technical guidelines'. The 'Service Guidelines and Standards for Water and Sanitation CCT (Vers 3 2)', 'guidelines' and 'Design Guidelines For Sewage Works Ontario Canada' each added 0 CQs, whereas the last file, '202-Technical guidelines 2004' added 1 to the total set.

The diminishing returns after a mere three standards is positive for the AgOCQs method in that domain experts do not have to spend days creating a large corpus, which otherwise would have cancelled out any gains in saving time authoring CQs manually.

[5] https://pypi.org/project/editdistance/

[6] https://huggingface.co/facebook/fasttext-language-identification

Table 1: Experiment 1 aggregate results from generation and human evaluation, of the two genres combined. (The symbol † denotes that at least one expert judged a question positively, or partially positive.)

Stage	|questions|	Pct.
Generation (all files)	11354	N/A
CLaRO filtering	2046	18
Similarity filtering	908	44
Selected for evaluation	200	
Ontologically acceptable	95	47
↳ guidelines (out of 100)		54
↳ papers (out of 100)		41
↳ Within scope†	58	29
↳ guidelines (out of 100)		39
↳ papers (out of 100)		28
↳ Of which relevant†	58	29
↳guidelines (out of 58)	28	48
↳ papers (out of 58)	30	52

3.2 Experiment 1: Results and Discussion of AgOCQs+

The pipeline generated 11354 initial questions of which 908 remained, as summarised in Table 1.

The human judgements on whether the questions were within scope averaged to 23% and of those judged within scope by at least one expert, 21% were deemed completely or partially relevant (26% of the full set of 200), as further summarised in Table 1. Examples of questions that were within scope, relevant for SewerNet, and of good quality as CQ for ontologies are included in Figure 2.

Overall, inter-annotator agreement was computed to be substantial (0.65) for scope and moderate (0.5) for relevance. Thus, there was no substantial difference by genre regarding scope and relevance, therewith falsifying H1 and H2.

The domain experts were surprised by the number of acronyms and abbreviations used in the questions and had to resort to the Web to check whether some were indeed within scope. Scope (i.e., related to sewer/wastewater or stormwater networks), was easier to evaluate than relevance. For instance, "What will the SSAIM contain?" in the evaluation set: SSAIM means Smart Sewer Asset Information Model, which was considered within scope and relevant. It was flagged as a problematic CQ on quality, however, principally because of the future tense.

Regarding quality of the questions, overall, about half (53%) were deemed problematic. The ratio of problematic questions was slightly higher for scientific papers (59%) while it was lower for standards (46%). Recurring issues included grammar (n=14), involving or asking for instances (n=48) rather than type-level knowledge, and content suitable for conceptual models rather than ontologies (n=23). For instance, "What is the name of the site?" and "What is the time taken to transverse the network?" are questions but not good as CQs for an ontology concerned with application independent knowledge. While it is not a high percentage, recent assessment of the CQ dataset that CLaRO was developed from was evaluated to have 23% problematic questions (Khan and Keet, 2024). That is, human authoring also faces quality issues, and this has an effect on CLaRO, and therewith the CLaRO filtering step.

- What is the rated capacity of the sewage treatment plant?
- What does the rainfall reduction method involve?
- What is the purpose of a diffuser?
- What is the purpose of an energy efficient treatment process?
- What is the purpose of a storm sewer system?
- What is the purpose of a major drainage system?
- What is the purpose of the two wastewater cycles?
- What is the definition of the pipe network?
- What is the transmission of Qs?
- What is the minimum height of the weir plate?

Figure 2: Sampling of CQs that were evaluated as within scope, relevant for SewerNet, and of acceptable quality in Experiment 1 (see supplementary material for a complete list).

3.3 Experiment 2: Results for AgOCQs++

The pipeline, with the aforementioned permutations and changes, initially generated 11330 questions and were eventually reduced to 2738 sentence, as summarised in Table 2.

In the final evaluation, an average of 23% of the questions were judged to be within scope and of the questions judged positively, there was an average of 69% questions judged to be relevant. A sampling of questions deemed in scope, relevant for the SewerNet ontology, and not considered problematic as CQ for an ontology is included in Figure 3. When analysing the expert annotations of the evaluated questions, we found that 32% still had grammar issues, which is worse than the 9% in Experiment 1. The agreement between the two experts was lower vs. Experiment 1 but it was still moderate (0.4) for both scope and relevance.

Table 2: Experiment 2 aggregate results from generation and human evaluation. (The symbol † denotes that at least one expert judged a question positively, or partially positive.)

Stage	|questions|	Pct.
Generation (all files)	11330	N/A
CLaRO filtering	7521	66
Similarity filtering	3510	47
Conceptual filter	2874	82
Non-English filter	2738	95
Selected for evaluation	48	
Ontologically acceptable	19	40
guidelines (out of 24)	6	25
papers (out of 24)	13	54
↳ Within scope†	16	33
↳ guidelines (out of 24)	6	25
↳ papers (out of 24)	10	42
↳ Of which relevant†	16	33
↳ guidelines (out of 16)	6	38
↳ papers (out of 16)	10	63

- During dry weather periods, what is the average daily flow of approximately m s?
- What is the minimum number of conduits connecting any manhole to the ground?
- What is the purpose of the proposed SSWMS?
- What should the valve and body be?

Figure 3: Sampling of CQs that were evaluated as within scope, relevant for SewerNet, and of acceptable quality in Experiment 2 (see supplementary material for a complete list).

The removal of contexts/inputs that are ungrammatical (extension E0) affected 9 'sentences' from the scientific papers and 15 'sentences' from the standards. As an example, the context "Huber L A Rossman R E Dickinson V P Singh D K Fervert Eds EPAStorm Water Management Model SWMM Chapter in Watershed Models CRC Press Boca Raton FL ISBN ISBN" was removed from the scientific articles and "DefinitionS Ventilated Improved Pit Toilet VIP toilet is a toilet which comprises..." was removed from the standards.

The grammar correction (extension E1) affected 3227 sentences of the 11330 total generated questions across the two data sets (747 for scientific papers and 2480 for standards). It corrected small typographical errors, such as from "... all of the activites?" to "... all of the activities?", grammar, such as correcting "... what is the charge of Irish Water?" into "...what is the charge for Irish Water?", and foreign language, such as from the generated question in German "Wo Wollen Sie sich fÃ¼r die Frage nach dem Grundstoff?" to have translates it into English as "Where will you go for the question after the basic substance?". The final filtering on English (extension E3) reduces the number of spurious foreign language sentences further, such as removing "Aktuelles und Hintergrundtextes bei uns?" that the trained model had generated from the input fragment "Standard Details Irish Water has developed Standard D etails describing typical infrastructure associated with the Works".

The conceptual filter removed questions such as "A What is the name of the company that has no AGB?" that are problematic as CQs for ontologies because they ask for an individual and a property (name) relevant in conceptual data modelling rather than for ontologies. Questions such as "What is an example of an existing utility?" were also removed, which may be borderline, as in some cases 'example of' seem more intended to ask for subclasses than individuals. Extension E2 did affect the results as follows. If it were to have been applied to the evaluated CQs of Experiment 1, then the scope percentage improves to 80%, relevance to 72%, and the percentage of unproblematic, i.e., possibly good CQs for ontologies, to 41%. For Experiment 2 with the revised pipeline, this 'conceptual filter' removed 222 sentences from those generated from the scientific articles and 414 based on the standards. Thus, the effects of the 'conceptual filter' was removal of 18% of the candidate CQ set fed to the filter.

Further, E3 affected 4% of the sentences. It was able to filter out questions that were completely of low quality for this task (e.g., "Um, is das KÃ¼rbis-Vehicles not beigemnt?"). Multilingual sentences presented a challenge since the non-English text could be interpreted as referring to a proper noun. For instance, the question "Vermittlungs-und Hybrid-Clubs. What type of services are available?" was not removed.

Finally, permutation P1 on filtering SQuAD on the CLaRO templates and training on the reduced set generated better output in the first step, but it had no effect for the final output, as the CLaRO filtering of the output equalised it (results not included).

4 Discussion

The data showed that AgOCQs+ with additional conceptual filtering and grammar correction, i.e., AgOCQs++, yielded the best results.

4.1 Answering the Research Questions

Regarding the specific questions from the introduction, the following. On RQ-1, i.e., whether the AgOCQs pipeline is also effective for use cases in other domains than it was initially tested on (COVID-19), it has been shown with the human evaluation by domain experts that it is somewhat effective for the domain of wastewater and stormwater networks as well, but also leaves room for improvement of the pipeline, and to aim for measures to increase the within-scope percentage in particular. Importantly, the whole AgOCQs+ and AgOCQs++ pipelines are now fully automated, simplifying and lowering the barrier to CQ generation for any other subject domain, and for reproducibility.

AgOCQs+ and AgOCQs++ are clearly effective on types of documents other than scientific articles, and possibly better, as shown in Experiment 1 (answering RQ-2). The effect of different corpus size on the number of CQs generated (RQ-3) showed that a small corpus already can generate a large number of relevant good quality CQs, and diminishing returns start at around the 5th document, as shown in the pre-experiment. The contribution of each filtering step on AgOCQs's output (RQ-4) and the effect of the SQuAD training set on the quality of the output (RQ-5) is discussed below.

While the average ratio of questions that are determined to be in scope is the same across the two evaluations, the ratio of relevant questions is higher for AgOCQs++. There is also a notable increase in the diversity of question structures. In Experiment 1, of the 200 questions that were evaluated, 193 of them fit one of the following patterns:

1. What is the purpose of ... (n=81)
2. What is the name of the ... (n=35)
3. What is an example of a ... (n=15)
4. What is the ... (n=47)
5. What will the ... (n=2)
6. What are the ... (n=6)
7. Who is the ... (n=2)
8. What does ... (n=3)
9. What are two ... (n=2)

In contrast, with respect to the 48 questions evaluated in Experiment 2, only 22 questions use the following patterns: "What is the ..." (n=18), "We are pleased to ..." (n=2), and "What was the ..." (n=2), and the rest of the questions, which make up 54% of the dataset, each have a distinct prefix and no obvious structural similarities.

The lack of diversity in question sentence structure is due to T5 and SQuAD and the issue does not appear to be easily corrigible by a range of strategies. While loosening the similarity to CLaRO patterns when filtering leads to increased diversity, it also increases the number of questions that include non-English text.

4.2 AgOCQs++ Pipeline Considerations

There are several other recurring issues. First, there are statements appended with a question mark, but grammatically they are not questions, and thus the pipeline has learned bad practice. For instance, while the pipeline generated questions such as "A list of the most common questions about the use of a scour chamber?" in Experiment 1, such questions were filtered out since their abstract form (i.e., "EC1 of EC2 about EC3 of EC4?" were not found in CLaRO. Such a strict similarity-CLaRO filter came at the expense of diversity in the final questions; hence, when it is loosened, the filter allows the generation of statements appended with a question mark to be presented as 'questions' (e.g., "Solicitation of construction and installation information in ADV?").

Second, the SQuAD questions come from, and are designed as a data set for, QA systems, and the questions are simple information-seeking and educational questions of the 'What is...' variety, which is narrower than the structures of the sentences for CQs for ontologies. T5 being trained

with the narrower set, it will then also much less likely generate more varied questions, as it repeats what it is trained on.

Third, in the question generation, it takes a subset of the paragraph. Supposedly it takes into account the context, i.e., the whole paragraph, but, based on our analysis, that is not what it is doing. T5 then produces out of context questions, often resorting to German and generating questions or statements either fully or partially in German. Similarly, if it selects a fragment that happens to have a formula or other generic text or a citation, it will generate an unrelated general domain question. We did not consider resolving this problem as it appears to be a problem with T5. Alternatively, one could pre-process the mini-corpus by cleaning it of strings that do not form part of a sentence, but this has the downside of additional time-consuming manual work.

There is no dataset available to train an LLM on generating questions from paragraphs of text, other than creating one from scratch specifically for CQs for ontologies. Also for few-shot prompting techniques as an alternative approach, many examples will have to be devised considering that ClaRo v2 has about 150 templates and an LLM would need several examples for each.

Notwithstanding these issues, the procedure does generate viable CQs for ontologies automatically that are traceable to the source. It also spurred further analysis into language characteristics of CQs, which may further contribute to language resources for ontology-related tasks.

5 Conclusions

We have demonstrated that AgOCQs++, now a fully automated pipeline, can generate competency questions that have the highest reported rate of being in scope and relevant, as judged by domain experts, where about half of the questions were deemed acceptable as competency questions for ontologies. The pipeline can generate questions for different genres of corpora, being at least scientific articles and guidelines and standards, with no significant difference in quality with respect to scope and relevance.

Future work will focus on creating a dataset of contexts and competency questions to alleviate the issues that arise due to the usage of SQuAD in the pipeline. Further research into metrics for measuring competency question quality will also be of value.

Limitations

The main limitation of the experiments is that it was evaluated with only one domain. This is the case also for experiments in related work, and thus more generally a shortcoming in the current state of research in automating CQ generation with LLMs. We hope that the updates to AgOCQs, particularly by having made it fully automated, will facilitate scaling up experimentation and use.

Supplementary material

The Experiment data are available at `https://github.com/AdeebNqo/AgOCQs_Plus`.

Acknowledgments

The authors would like to thank Siyanda Makhathini for his contribution to the preliminary experimentation. This research has received support from the European Union's Horizon research and innovation program (under the MSCA-SE (Marie Skłodowska- Curie Actions Staff Exchange) grant agreement 101086252; Call: HORIZON-MSCA-2021-SE-01; Project title: STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management). This research has also received support from the ANR CROQUIS (Collecte, représentation, complétion, fusion et interrogation de données de réseaux d'eau urbains hétérogènes et incertaines) project, grant ANR-21-CE23-0004 of the French research funding agency - Agence Nationale de la Recherche (ANR).

References

Reham Alharbi, Valentina Tamma, Floriana Grasso, and Terry R. Payne. 2023. An experiment in retrofitting competency questions for existing ontologies. *ArXiv*, abs/2311.05662.

Reham Alharbi, Valentina Tamma, Floriana Grasso, and Terry R. Payne. 2024a. An experiment in retrofitting competency questions for existing ontologies. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC 2024*, pages 1650–1658. ACM. Avila, Spain, April 8-12, 2024.

Reham Alharbi, Valentina Tamma, Floriana Grasso, and Terry R. Payne. 2024b. A review and comparison of competency question engineering approaches. In *Knowledge Engineering and Knowledge Management - 24th International Conference, EKAW, 2024,*

Proceedings, volume 15370 of *LNCS*, pages 271–290. Springer. 2024, Amsterdam, The Netherlands, November 26-28.

Mary-Jane Antia and C. Maria Keet. 2021. Assessing and enhancing bottom-up CNL design for competency questions for ontologies. In *Proceedings of the Seventh International Workshop on Controlled Natural Language (CNL 2020/21)*, Amsterdam, Netherlands. Special Interest Group on Controlled Natural Language.

Mary-Jane Antia and C. Maria Keet. 2023. Automating the generation of competency questions for ontologies with agocqs. In *Knowledge Graphs and Semantic Web*, pages 213–227, Cham. Springer Nature Switzerland.

Camila Bezerra and Fred Freitas. 2017. Verifying description logic ontologies based on competency questions and unit testing. In *ONTOBRAS'17*, pages 159–164.

Camila Bezerra, Fred Freitas, and Filipe Santana. 2013. Evaluating ontologies with competency questions. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03*, WI-IAT '13, pages 284–285. IEEE Computer Society.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *(To appear)*, 7(1):411–420.

C. M. Keet, B. Haydar, and N. Chahinian. 2025. The sewernet domain ontology: on clarifying and harmonising terminology. In *EGU General Assembly 2025*, pages EGU25–8662. Vienna, Austria, 27 Apr–2 May 2025.

C. M. Keet and A. Lawrynowicz. 2016. Test-driven development of ontologies. In *Proceedings of the 13th Extended Semantic Web Conference (ESWC'16)*, volume 9678 of *LNCS*, pages 642–657, Berlin. Springer. 29 May - 2 June, 2016, Crete, Greece.

C. M. Keet, Z. Mahlaza, and M.-J. Antia. 2019. CLaRO: a controlled language for authoring competency questions. In *13th Metadata and Semantics Research Conference (MTSR'19)*, volume 1057 of *CCIS*, pages 3–15. Springer. 28-31 Oct 2019, Rome, Italy.

Zubeida Khan and C. Maria Keet. 2024. On the roles of competency questions in ontology engineering. In *24th International Conference on Knowledge Engineering and Knowledge Management (EKAW'24)*, volume 15370 of *LNAI*, pages 123–132. Springer. November 26-28, Amsterdam, The Netherlands.

Xueli Pan, Jacco van Ossenbruggen, Victor de Boer, and Zhisheng Huang. 2025. A rag approach for generating competency questions in ontology engineering. *Preprint*, arXiv:2409.08820.

Jedrzej Potoniec, Dawid Wisniewski, Agnieszka Lawrynowicz, and C. Maria Keet. 2020. Dataset of ontology competency questions to SPARQL-OWL queries translations. *Data in Brief*, 29:105098.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. Technical report.

Mari Carmen Suarez-Figueroa, Guadalupe Aguado de Cea, Carlos Buil, Klaas Dellschaft, Mariano Fernandez-Lopez, Andres Garcia, Asuncion Gómez-Pérez, German Herrero, Elena Montiel-Ponsoda, Marta Sabou, Boris Villazon-Terrazas, and Zheng Yufei. 2008. NeOn methodology for building contextualized ontology networks. NeOn Deliverable D5.4.1, NeOn Project.

Elodie Thiéblin, Ollivier Haemmerlé, and Cassia Trojahn. 2018. Complex matching based on competency questions for alignment: a first sketch. In *13th International Workshop on Ontology Matching (OM@ISWC 2018)*, pages 66–70, Monterey, US. CEUR-WS.

Mike Uschold and Michael Gruninger. 1996. Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–136.

Dawid Wisniewski, Jedrzej Potoniec, and Agnieszka Lawrynowicz. 2021. SeeQuery: An automatic method for recommending translations of ontology competency questions into SPARQL-OWL. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management*, pages 2119–2128. ACM. Virtual Event, Queensland, Australia, November 1 - 5, 2021.

Dawid Wisniewski, Jedrzej Potoniec, Agnieszka Lawrynowicz, and C. Maria Keet. 2019. Analysis of ontology competency questions and their formalisations in sparql-owl. *Journal of Web Semantics*, 59(100534):19p.

A Appendix A: The mini-corpora

Guideline documents:

- Code of Practice for Wastewater Infrastructure. Connections and Developer Services: Design and Construction Requirements for Self-Lay Developments. July 2020 (Revision 2), Document IW-CDS-5030-03. https://www.water.ie/sites/default/files/docs/connections/faqs/Wastewater-Code-of-Practice.pdf

- Service Guidelines & Standards, Water and Sanitation Department, City of Cape Town. 20 June 2019 (Version 3.3) https://resource.capetown.gov.za/documentcentre/Documents/Procedures,%20guidelines%20and%20regulations/Service%20Guidelines%20and%20Standards_for_Water_and_Sanitation_CCT%20(Vers%203%202).pdf

- Design Guidelines For Sewage Works, Ontario Ministry of the Environment Sewage Technical Working Group, Hydromantis, Inc., and XCG Consultants Ltd. ISBN 978-1-4249-8438-1. PIBS 6879. https://www.ontario.ca/document/design-guidelines-sewage-works-0

- Technical guidelines for the development of water and sanitation infrastructure, Department of Water Affairs and Forestry. Second Edition: 2004. https://www.fsmtoolbox.com/assets/pdf/202_-_Technical_guidelines_2004.pdf

Scientific papers:

- C. Montalvo, J.D. Reyes-Silva, E. Sañudo, L. Cea, J. Puertas. Urban pluvial flood modelling in the absence of sewer drainage network data: A physics-based approach. Journal of Hydrology, 634, 2024, 131043. DOI: 10.1016/j.jhydrol.2024.131043.

- Gabriel Perez, Jesus D. Gomez-Velez, Stanley B. Grant, The sanitary sewer unit hydrograph model: A comprehensive tool for wastewater flow modeling and inflow-infiltration simulations. Water Research, 249, 2024, 120997. DOI: 10.1016/j.watres.2023.120997.

- Vikki Edmondson, Martin Cerny, Michael Lim, Barry Gledson, Steven Lockley, John Woodward. A smart sewer asset information model to enable an 'Internet of Things' for operational wastewater management. Automation in Construction, 91, 2018, 193-205. DOI: 10.1016/j.autcon.2018.03.003.

- Priyan Malarvizhi Kumar, Choong Seon Hong. Internet of things for secure surveillance for sewage wastewater treatment systems. Environmental Research, 203, 2022, 111899, DOI: 10.1016/j.envres.2021.111899.