

Integrating Domain Experts and Large Language Models for Competency Question Generation

Maryam Sefidbari^{1,*}, Leyla Jael Castro¹

¹ZB MED – Information Center for Life Sciences, Cologne, Germany

Abstract

Competency Questions (CQs) play a central role in Ontology Engineering (OE) by capturing domain requirements. However, manual CQ elicitation is time-consuming and requires domain expertise. We present a multi-stage framework integrating domain expert input and LLMs for semi-automated generation of CQs. The framework leverages Azure OpenAI Service for CQ generation and processing. Beginning with a lightweight expert scoping step, it automatically constructs a domain corpus from a set of scholarly articles, generates and refines CQs, removes redundancy, scores relevance, and organizes the resulting CQs into thematic clusters. We evaluate the framework within two different domains: *Building damage and construction defect analysis*, and *Proton Exchange Membrane Fuel Cell degradation mechanisms*, and across three dimensions: intrinsic quality assessed by independent domain experts, comparative performance against manually authored gold-standard CQs, and coverage analysis quantifying novelty versus reproduction. Results show that generated CQs achieve high clarity, atomicity and domain relevance with strong inter-rater agreement, and adequately reproduce 90.2% of expert-defined requirements while also producing novel CQs, making LLMs an effective complement to human expertise in OE.

Keywords

Competency Questions, Ontology Engineering, Ontology Validation, Large Language Models, Requirements Elicitation

1. Introduction

Competency Questions (CQs) are used in Ontology Engineering (OE) to capture domain requirements and validate ontology adequacy [1] as they facilitate requirement specification and scope definition [1, 2]. However, CQ elicitation remains a bottleneck as manual elicitation is time-consuming [1, 3] and structured approaches, e.g., CORAL [4] and CLaRO [3], remain constrained by template dependency, requiring CQs to be authored or reformulated according to fixed lexico-syntactic patterns. Recent LLM-based approaches [5, 6, 7] are promising but typically rely on single-step prompting strategies, insufficient to guarantee structural quality and systematic domain coverage. To address this gap, we propose a multi-stage framework for semi-automated CQ generation, integrating domain expert guidance (in the first phase) with LLM-based generation and refinement (in the next phases).

Starting with four domain-scoping questions, the pipeline extracts scholarly literature, generates CQs, refines them through entity abstraction, filters them through redundancy removal, relevance scoring, and linguistic complexity analysis, and organizes resulting CQs into thematic clusters. Expert involvement is deliberately front-loaded in the initial scoping step, and the remaining five phases run automatically without further expert intervention. The framework is evaluated in two domains: (1) building damage and construction defect analysis, and (2) Proton Exchange Membrane Fuel Cell (PEMFC) degradation mechanisms. Our main contributions are: (1) a human-LLM collaborative pipeline for semi-automated CQ generation grounded in domain literature; (2) a multi-stage refinement and filtering pipeline enhancing CQ generalizability and structural quality; and (3) a rigorous empirical evaluation demonstrating how the framework complements human expertise for CQ elicitation.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author. Email:sefidbari@zbmed.de

✉ sefidbari@zbmed.de (M. Sefidbari); ljgarcia@zbmed.de (L.J. Castro)

ORCID 0009-0008-7998-3965 (M. Sefidbari); 0000-0003-3986-0510 (L.J. Castro)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related Work

Competency Questions (CQs), introduced by Grüninger and Fox [1], are natural language questions an ontology must answer, serving as functional requirements that define ontology scope, identify key concepts, and provide validation criteria. High-quality CQs should be atomic, clear, domain-relevant, and testable [1, 3]. Traditionally, CQs are manually elicited from domain experts, a process that is both time-consuming and cognitively demanding [3]. To address these limitations, several approaches have been proposed to support or automate CQ generation.

Recently, LLM-based approaches have emerged, demonstrating that transformer models can generate plausible CQ candidates from domain descriptions with minimal input [5, 6]. Mahlaza et al. [5] introduced AgOCQs++, based on AgOCQs [8], taking a corpus-based approach using a fine-tuned T5 model with CLaRO-based filtering. Pan et al. [9] proposed a Retrieval-Augmented Generation approach grounding LLM generation in retrieved ontological content. NeOn-GPT [10] integrates LLMs into structured Ontology Engineering workflows using prompt templates, while Rebboud et al. [6] conducted a comparative study of six LLMs for CQ generation, both [10] and [6] remain single-step approaches without refinement mechanisms. Also, OntoChat [7] supports CQ generation through conversational prompting.

Similar to [8] and [9], we leverage domain literature as a primary knowledge source for CQ generation. Same as [7], we aim to streamline stakeholder interaction through domain expert input. We extend these approaches by combining expert guided scoping with automated scholarly article retrieval, LLM-based refinement, redundancy filtering, relevance scoring, linguistic complexity analysis, and semantic clustering within a single cohesive pipeline. We also provide a rigorous empirical evaluation across three dimensions: intrinsic quality of generated CQs, comparative performance against gold-standard CQs, and reproduction vs. novelty analysis against gold-standard dataset.

3. Framework

Our proposed framework uses *gpt-4o-mini* and comprises a 6-phase pipeline as illustrated in Figure 1.

Phase 1 — Domain Information Extraction. The domain expert responds to four questions defining the ontology scope, purpose, intended users, and use-cases. These responses are then processed by LLM to extract structured domain parameters including a main domain name of 5-10 words representing the core area; 20-30 compound topic terms suitable for article search queries; 30-50 single filter keywords directly relevant to the domain; 2 core compound general terms used to anchor search queries; 3-5 single main domain words; and 5-10 high-level ontology coverage areas. The specified quantities were established through iterative testing of the pipeline. We observed that adhering to these ranges consistently yielded more relevant scholarly articles and, consequently, higher-quality CQs. Given the critical influence of this phase on the scope and relevance of CQs, we recommend an expert validation step before proceeding to article extraction, thereby safeguarding the semantic alignment of the entire process.

Phase 2 — Scholarly Article Extraction. This phase constructs a comprehensive domain-specific corpus from scholarly literature, which serves as the foundational knowledge base for CQ generation. The process consists of four sub-stages: query construction and execution, multi-stage article filtering, corpus finalization, and snippet extraction. Search queries combining topic terms and general domain terms are automatically created and run over Google Scholar via *SerpAPI*. Up to 31 unique queries yield up to 1,240 article references. Next, the articles undergo multi-stage filtering, selecting articles in English, with more than 1,500 characters, within the last 15 years, and containing at least three core domain words. Further keyword filtering verifies the presence of at least 5 filter keywords, plus lemmatized topic term matching, which requires all lemmas of at least 3 topic terms in the text.

Filtering continues until all 20-30 topic terms are covered or a maximum of 30 articles are collected (due to the token budget). To create a balanced input for the LLM, relevant text snippets are extracted from the finalized corpus using a round-robin sampling algorithm that selects up to 3 snippets per article.

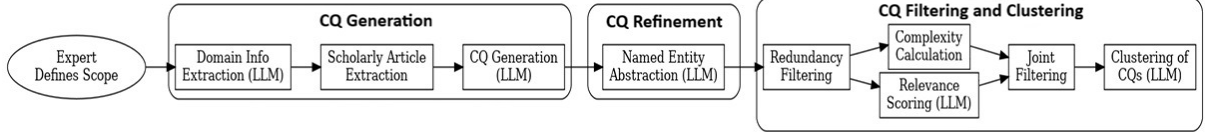


Figure 1: The proposed LLM-powered framework for CQ generation.

Snippets are defined as paragraphs containing at least one filter keyword. A global token budget of 50k tokens is enforced to respect computational limits while maintaining diverse corpus representation.

Scholarly articles were selected as the foundational corpus given the technical and research-active nature of our two evaluation domains. Other domains may require alternative or complementary sources, which we leave for future work. Also, all numerical thresholds reported above are configurable parameters of the framework; the values given here reflect the specific instantiation used in our evaluation in the domains (Section 4).

Phase 3 – CQ Generation. The LLM acts as a *Senior Ontology Engineer* using few-shot [11] and Chain-of-Thought [12] prompting techniques. At temperature=0.3, it synthesizes article snippets to generate a minimum of 80 domain-relevant and atomic CQs covering all predefined ontology coverage areas. An atomic CQ expresses a single, indivisible information requirement. An initial attempt with temperature=0.1 produced highly repetitive CQs with limited lexical and structural variation, while temperature=0.3 balanced output diversity with adherence to the required atomic CQ structure.

Phase 4 – CQ Refinement. Based on the OntoChat [7] approach, generated CQs undergo a named entity abstraction. The LLM identifies and replaces specific real-world named entities – such as materials, technologies, locations – with generalized placeholder terms. This step enhances the ontological generalizability and formalizability of CQs and prevents them from being overfitted to individual examples drawn from retrieved articles.

Phase 5 – CQ Filtering. The filtering phase eliminates redundant, irrelevant, and complex CQs. Refined CQs are de-duplicated using a hybrid similarity score combining semantic similarity from the `a11-mpnet-base-v2` sentence transformer [13] (weight: 0.7) and syntactic similarity from the Jaccard index (weight: 0.3). Questions above a redundancy threshold of 0.75 are clustered into connected components within a similarity graph using depth-first search. From each cluster, a single representative question is selected, by default the shortest. A threshold of 0.75 was selected empirically to balance aggressive de-duplication against the risk of merging distinct but topically related CQs; lower thresholds collapsed semantically distinct questions, while higher thresholds retained near-duplicate phrasings.

Next, unique CQs undergo relevance scoring by LLM, evaluating alignment with the expert-defined ontology scope on a 4-point Likert scale: (4) explicitly matches requirements (3) implicit but clearly inferable requirement (2) tangentially related (1) irrelevant. Integer relevance values undergo non-linear normalization ($1 \rightarrow 0.0$, $2 \rightarrow 0.5$, $3 \rightarrow 0.75$, $4 \rightarrow 1.0$) amplifying the distinction between high-relevance questions while sharply penalizing marginal or irrelevant ones. Then, each CQ undergoes linguistic complexity analysis using spaCy’s dependency parsing. The complexity metric quantifies structural intricacy through a weighted combination of five linguistic features: noun phrases ($w=0.3$), verbs ($w=0.2$), prepositions ($w=0.15$), conjunctions ($w=0.15$), and modifiers ($w=0.2$), normalized by a factor of 5.0. The raw complexity score is normalized to $[0,1]$, with higher values indicating more complex, wordy questions. Finally, a joint filtering function combines both scores:

$$final_score = 0.8 * normalized_relevance + 0.2 * (1 - complexity).$$

Questions are retained if their final score exceeds a threshold of 0.5. The weighting scheme deliberately prioritizes domain relevance (0.8) as the dominant criterion, reflecting its greater importance for OE, while linguistic complexity (0.2) serves as a secondary quality control mechanism that penalizes overly wordy or structurally compound questions.

Phase 6 – CQ Clustering. Finally, filtered CQs are organized into 5–7 thematic clusters by LLM. Each cluster is assigned a meaningful title representing a coherent knowledge area, revealing the domain’s conceptual structure. The complete source code implementing the framework, expert responses to four

scoping questions, retrieved scholarly articles, gold-standard dataset, and generated CQs are publicly available at <https://github.com/Maryam-sfb/CQ-Generation-Framework> for transparency and reuse.

4. Evaluation

The framework was evaluated across three dimensions within two different domains. The domains were selected to test the framework under contrasting conditions: building damage and construction defect analysis is a practically-oriented field with fragmented, heterogeneous literature, while PEMFC degradation mechanisms is a mature, well-studied technical domain with an existing gold-standard CQ set (DFC ontology¹) available for benchmarking.

Intrinsic Quality Assessment. In a single execution on the domain of *Building Damage and Construction Defect Analysis*, 80 CQs were generated, reduced to a final set of 43 after redundancy removal and joint filtering. Two domain experts with 15 and 17 years of experience² independently evaluated the CQs using identical 4-point Likert scales across three dimensions: Domain Relevance (from irrelevant to highly relevant), Atomicity (highly composite to perfectly atomic), and Clarity (confusing to extremely clear). Inter-rater reliability was assessed using Weighted Kappa with quadratic weights, using established thresholds: <0.40 (Fair), $0.40-0.60$ (Moderate), $0.60-0.80$ (Substantial), ≥ 0.80 (Almost Perfect), alongside ICC(3,1) for completeness. In addition, the amount of coverage of the generated CQs over the ontology coverage areas was assessed by one of the domain experts, who classified each CQ into one or more of 5 predefined ontology coverage areas.

Table 1 presents inter-rater reliability and the consensus quality scores for LLM-generated CQs. Coverage analysis revealed that all five ontology coverage areas are represented with 15 to 24 CQs per area (balance ratio: 0.625), and 69.8% of CQs span multiple coverage areas simultaneously, indicating strong conceptual integration across the domain knowledge structure.

Table 1

Intrinsic Quality Evaluation Results (N=43 CQs). κ : Weighted Cohen’s Kappa; Exact Agr.: percentage of CQs with identical scores from both raters; Class.: qualitative reliability classification based on κ .

Metric	Mean $\pm$ SD	κ	Exact Agr.	Class.
Clarity	4.00 $\pm$ 0.00	1.000	100%	Excellent
Atomicity	3.93 $\pm$ 0.23	0.645	95.3%	Excellent
Domain Relevance	3.45 $\pm$ 0.48	0.700	83.7%	Good
Overall	3.79	—	—	Excellent

Comparative Performance Against Manual Elicitation. The framework was benchmarked against 41 manually authored gold-standard CQs from the DFC ontology in the domain of PEMFC degradation mechanisms. From a single pipeline execution, 68 CQs remained after joint filtering, yielding a combined dataset of 109 CQs. A blind evaluation was conducted using Claude Sonnet 4.5, on all 109 CQs in shuffled order (random_state=42) on a 5-point Likert scale across four dimensions: Domain Relevance, Atomicity, Clarity, and Importance. Mann-Whitney U tests with Cohen’s d effect sizes were used for statistical comparison.

Table 2 presents the comparative results. The analysis reveals a complementary strength pattern between the two sources. Expert-authored CQs outperform generated CQs in Domain Relevance ($p = 0.0018$, $d = -0.63$, medium effect) and Importance ($p = 0.0024$, $d = -0.67$, medium effect), reflecting their stronger alignment with specific coverage areas and superior identification of critical ontological competencies. However, framework-generated CQs outperform expert CQs in Clarity ($p = 0.0068$, $d = 0.48$, small effect), with 91% of generated CQs receiving the maximum score compared to 73% of expert CQs. Atomicity shows no significant difference ($p = 0.0693$), although generated CQs achieve perfect consistency (5.00 ± 0.00) compared to near-perfect expert scores (4.95 ± 0.22). Finally, domain

¹The Durafuelcell project, Future Engineering Group, Technische Hochschule Nürnberg, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 528480942 – FIP8.

²Many thanks to Amir Haji Mohamadi and Morteza Sanaei for their valuable contribution throughout this evaluation phase.

terminology analysis confirms equivalent vocabulary incorporation between two sources ($p = 0.5589$, $d = 0.13$), with LLM-generated CQs averaging 3.90 domain terms per question compared to 3.68 for expert-authored CQs, validating the framework’s ability to learn and apply domain-specific terminology from the retrieved literature.

Table 2

Quality Metrics: Expert-Authored vs. Framework-Generated CQs. p : Mann-Whitney U test p -value; d : Cohen’s d effect size; Effect: qualitative interpretation of d .

Metric	Expert	Generated	p	d	Effect
Relevance	4.34±0.66	3.96±0.58	0.0018	−0.63	Medium
Atomicity	4.95±0.22	5.00±0.00	0.0693	0.37	Small
Clarity	4.71±0.46	4.90±0.35	0.0068	0.48	Small
Importance	3.56±0.90	3.04±0.68	0.0024	−0.67	Medium

Reproduction vs. Novelty Analysis. The third evaluation phase examines whether the framework primarily reproduces requirements captured in human-authored CQs or generates novel questions. A semantic coverage analysis was conducted comparing 80 LLM-generated CQs against the same 41 human-authored CQs of the DFC ontology in the PEMFC domain. Claude Sonnet 4.5 performed a systematic four-step semantic assessment for each of the 41 human-authored CQs: (1) interpreting the question to identify its core knowledge requirement including relevant concepts, relationships, and scope; (2) comparing it against all 80 LLM-generated CQs to identify all semantically relevant matches; (3) assigning a coverage percentage (0–100%) based on semantic equivalence, considering both explicit overlap and implicit fulfilment of the knowledge requirement; and (4) validating consistency by reviewing scores across similar CQ types.

The framework achieved 59.4% average coverage (median: 65%) with 90.2% adequate coverage ($\geq 30\%$) of human-authored CQs, indicating that most of the expert-defined requirements are partially addressed by LLM-generated CQs. Only 9.8% of human-defined CQs were completely uncovered, while 34.1% achieved strong reproduction ($\geq 80\%$). Table3 presents the overall coverage statistics.

Table 3

Overall Semantic Coverage Statistics

Metric	Value	Interpretation
Average Coverage	59.4%	Moderate reproduction with substantial novelty
Adequate Coverage ($\geq 30\%$)	90.2% (37/41)	Most human requirements are adequately addressed
Partial Coverage (30–79%)	48.8% (20/41)	Nearly half of requirements are partially covered
Strong Coverage ($\geq 80\%$)	34.1% (14/41)	One-third of requirements are strongly covered
Complete Gaps (0%)	9.8% (4/41)	Small number of uncovered requirements

Furthermore, category-wise analysis of generated CQs reveals that core topics such as catalyst degradation (87%) and dynamic loading (85%) achieve high coverage, reflecting their abundant representation in the retrieved literature. On the other hand, specialized topics such as economics and domain-specific measurement technologies show complete gaps (0%), highlighting the importance of expert review and validation for comprehensive coverage of niche knowledge areas.

5. Conclusion and Future Work

We presented a multi-stage framework integrating domain expert input with LLMs for semi-automated CQ generation. By combining a lightweight expert scoping with automated literature retrieval, CQ generation, refinement, redundancy filtering, relevance scoring, and thematic clustering, the framework reduces reliance on continuous expert involvement while producing structurally sound and generalizable CQs. Our evaluation across two domains showed that generated CQs achieve high atomicity, clarity and domain relevance, with strong inter-rater agreement, complementing human expertise in requirements

elicitation. Also, coverage analysis revealed that 90.2% of expert-defined knowledge requirements are adequately addressed by the framework-generated CQ set.

Future work will focus on LLM variations, favoring open models whenever possible, customization of the criteria used to select scholarly articles, cross-domain validation across other knowledge domains, and longitudinal studies examining the framework's practical impact on ontology quality and development efficiency. An interactive web interface allowing domain experts to monitor CQ generation at each stage, enabling expert checkpoints at later pipeline stages and validate the final output is currently under development.

Declaration on Generative AI

Claude Sonnet 5 was used to refine the sentences. The authors reviewed and revised all AI-generated content and are fully responsible for the content of this publication.

References

- [1] M. Grüninger, M. S. Fox, The role of competency questions in enterprise engineering, in: *Benchmarking – Theory and Practice*, Springer, 1995, pp. 22–31.
- [2] N. F. Noy, D. L. McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*, Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, 2001.
- [3] Y. Ren, A. Parvizi, C. Mellish, et al., Towards competency question-driven ontology authoring, in: *The Semantic Web: Trends and Challenges – 11th International Conference, ESWC 2014*, Springer, 2014, pp. 752–767. doi:10.1007/978-3-319-07443-6_50.
- [4] A. Fernández-Izquierdo, M. Poveda-Villalón, R. García-Castro, CORAL: A corpus of ontological requirements annotated with lexico-syntactic patterns, in: *The Semantic Web – ESWC 2019*, Springer, 2019, pp. 443–458. doi:10.1007/978-3-030-21348-0_29.
- [5] Z. Mahlaza, C. M. Keet, N. Chahinian, B. Haydar, On the feasibility of LLM-based automated generation and filtering of competency questions for ontologies, in: *Proceedings of the 5th Conference on Language, Data and Knowledge (LDK 2025)*, 2025, pp. 136–146.
- [6] Y. Rebboud, L. Tailhardat, P. Lisena, R. Troncy, Can LLMs generate competency questions?, in: *The Semantic Web: ESWC 2024 Satellite Events*, Springer, 2024, pp. 71–80.
- [7] B. Zhang, V. A. Carriero, K. Schreiberhuber, et al., OntoChat: A framework for conversational ontology engineering using language models, in: *The Semantic Web: ESWC 2024 Satellite Events*, Springer, 2024, pp. 102–121.
- [8] M. J. Antia, C. M. Keet, Automating the generation of competency questions for ontologies with AgOCQs, in: *Iberoamerican Knowledge Graphs and Semantic Web Conference*, Springer, 2023, pp. 213–227.
- [9] X. Pan, J. van Ossenbruggen, V. de Boer, Z. Huang, A RAG approach for generating competency questions in ontology engineering, in: *Metadata and Semantic Research – 18th Research Conference, MTSR 2024*, Springer, 2025, pp. 70–81. doi:10.1007/978-3-031-81974-2_6.
- [10] N. Fathallah, A. Das, S. D. Giorgis, et al., NeOn-GPT: A large language model-powered pipeline for ontology learning, in: *The Semantic Web: ESWC 2024 Satellite Events*, Springer, 2024, pp. 36–50.
- [11] T. B. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, in: *Advances in Neural Information Processing Systems*, volume 33, 2020, pp. 1877–1901.
- [12] J. Wei, X. Wang, D. Schuurmans, et al., Chain-of-thought prompting elicits reasoning in large language models, in: *Advances in Neural Information Processing Systems*, volume 35, 2022, pp. 24824–24837.
- [13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, 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)*, Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.