Comparing Methods for Competency Question Elicitation from Ontology Requirements

Reham Alharbi[1], Jacopo de Berardinis[1], Terry R. Payne[1] and Valentina Tamma[1]

[1]University of Liverpool, Liverpool, UK

Abstract

Competency Questions (CQs) are used guide ontology development, yet formulating them in such a way as to align them to the stakeholder needs remains challenging. This paper presents a comparative analysis of three CQ elicitation methods: manual authoring by ontology engineers; template-based instantiation; and automated generation using different LLMs (GPT-4.1, Gemini 2.5). Each CQ is evaluated across dimensions of suitability, readability, and complexity. To facilitate this evaluation we introduce **AskCQ**, a dataset of 204 CQs derived from a shared user story in the cultural heritage domain. Our results show that manually authored CQs are consistently more acceptable, readable, and concise. LLM-generated CQs are more complex and diverse but require refinement. These findings highlight the importance of human expertise and suggest potential hybrid approaches.

1. Introduction

Competency Questions (CQs) are a foundational tool in ontology engineering (OE), used to scope knowledge models, validate implementations, and align ontologies with stakeholder needs. Despite their value, CQs are often under used in practice due to their perceived complexity, lack of tooling, and absence of clear evaluation criteria [1, 2]. Although various automated and semi-automated CQ generation methods have emerged (of which several use templates or Large Language Models (LLMs)) [3], there are few studies to date that have addressed the systematic evaluation of their performance with respect to the characteristics of the CQ that they generate. As a result, practitioners struggle to create questions that are clear, consistent, and aligned with intended requirements.

To lower the barrier of CQ adoption, a number of semi-automated and automated generation methods have been proposed, that range from template-based instantiations [4] to generative language models [5, 6]. However, prior work has rarely compared the outputs of these methods side by side using a shared set of ontology requirements. Therefore, our study addresses this gap by evaluating three CQ elicitation approaches: *manual*; *pattern-based*; and *LLM-based*, on a shared scenario from the cultural heritage domain. We introduce **AskCQ**, a novel multi-annotator CQ dataset derived from the same user story using different elicitation methods. The study evaluates the generated CQs based several criteria: expert-rated suitability; ambiguity; readability; relevance to the user story: and linguistic/ontological complexity. We also assess semantic similarity across the different CQ sets to quantify overlap and novelty. Our findings provide actionable insight into the trade-offs and biases introduced by the different CQ elicitation approaches. The dataset and code used are available at: https://github.com/KE-UniLiv/askcq

2. Methodology

To investigate the impact of different Competency Question (CQ) elicitation strategies on the characteristics of the resulting questions, we conducted a comparative analysis across three representative

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

R.Alharbi@liverpool.ac.uk (R. Alharbi); Jacopo.De-Berardinis@liverpool.ac.uk (J. de Berardinis); T.R.Payne@liverpool.ac.uk (T. R. Payne); V.Tamma@liverpool.ac.uk (V. Tamma)

0000-0002-8332-3803 (R. Alharbi); 0000-0001-6770-1969 (J. de Berardinis); 0000-0002-0106-8731 (T. R. Payne); 0000-0002-1320-610X (V. Tamma)

approaches: fully manual authoring; template-based instantiation; and automatic generation via Large Language Models (LLMs). Each approach was applied independently to the same ontology requirement source to ensure a fair and controlled comparison.

1. **Manual (Human-Authored):** Two ontology engineers (HA-1 and HA-2), each with over five years of professional experience in ontology design and requirement engineering, independently read and interpreted the same user story. Based solely on their expert understanding of the personas, goals, and informational needs described therein, each formulated a set of CQs without constraints on format or style. This condition serves as the expert-driven baseline and reflects common manual practice in ontology development.
2. **Template-Based (Pattern Instantiation):** An ontology engineer with similar domain experience instantiated a curated set of 19 CQ patterns derived from Ren et al. [4]. These patterns use archetypal structures such as "Which [CE1] [OPE] [CE2]?" and "Is the [CE1] [CE2]?" and were manually populated with entities and relations extracted from the user story. The instantiation process required the identification of suitable fillers from the story content, and their mapping to the syntactic slots defined by the patterns. This semi-automated method offers structured linguistic support but limited flexibility.
3. **LLM-Based (Generative AI):** Two state-of-the-art LLMs — GPT-4.1 and Gemini 2.5 Pro — were prompted to generate CQs directly from a markdown-formatted version of the user story. Prompts were intentionally minimal and neutral: no explicit instructions were given regarding CQ format, number, or examples, to avoid priming or biasing the output. This open-ended configuration was intended to test each model's intrinsic ability to extract ontology-relevant requirements and phrase them as competency questions.

2.1. AskCQ Dataset Construction

All three approaches were applied to the same textual requirement source: a detailed *user story* developed for a cultural heritage ontology use case. The story, adapted from the methodology proposed by de Berardinis et. al. [7] is centred on two personas, a music archivist and a curator, and describes their activities and data needs relating to a museum's music memorabilia collection, including acquisition, loan, metadata management, and display. The output comprises five CQ sets: HA-1 (44 CQs), HA-2 (54 CQs), Pattern (38 CQs), GPT-4.1 (26 CQs), and Gemini 2.5 Pro (42 CQs), totalling 204 distinct questions.[1]

2.2. Evaluation Dimensions and Feature Extraction

To assess the quality and characteristics of the generated CQs, we adopted a multi-dimensional, mixed-methods evaluation framework encompassing both qualitative expert judgment and quantitative feature analysis: CQ suitability, structural and semantic properties, and inter-method agreement.

1. Suitability (Expert Evaluation): Each CQ was independently reviewed by three ontology experts, who rated its acceptability for guiding ontology engineering in the context of the user story. Scores ranged from -3 (unanimous rejection) to +3 (unanimous acceptance). The experts were not provided with explicit criteria to preserve their interpretive autonomy, analogous to the elicitation setup. A Fleiss' Kappa of $\kappa = 0.35$ indicated fair inter-expert agreement.

2. Readability: Each CQ was assessed to gauge its ease of understanding. We assess readability in a similar way to Ciroku, et al. [5], where a suite of established readability indices designed to capture different aspects of textual difficulty were initially computed for each CQ using the `textstat` Python

[1]The resulting **AskCQ** dataset is publicly released under a CC-BY license, and all CQs were anonymized and randomly shuffled prior to evaluation to minimize bias regarding their origin.

library.[2] In this paper we report only the Flesch-Kincaid Grade Level (FKGL) and the Dale-Chall Readability Score (DCR) as representative readability features.

- **Flesch-Kincaid Grade Level (FKGL)** — Estimates the education level (U.S. grade) required for comprehension [8].
- **Dale-Chall Readability Index (DCR)** — Penalizes complex vocabulary based on a restricted list of familiar words [9].

3. Relevance: The alignment of each CQ, together with the user story, was assessed by Gemini 2.5 Pro and rated on a 4-point scale Likert scale using the following criteria: (4) directly stated in the story, (3) inferable and necessary, (2) tangentially relevant, (1) off-topic. The evaluation prompt was carefully designed and spot-validated on a selected sample of CQs.

4. Complexity: The following four complementary metrics were defined to quantify different facets of CQ complexity:

- **c0 (Length):** The total number of characters, as a coarse indicator of verbosity and potential elaboration.
- **c1 (Requirement Complexity):** The number of distinct concepts, properties, relations, and filters identified in the CQ by Gemini 2.5 Pro.
- **c2 (Linguistic Complexity):** A count of syntactic and lexical features (noun phrases, verbs, prepositions, modifiers, etc.) extracted via `spaCy`.
- **c3 (Syntactic Complexity):** The structural depth and richness of the dependency parse tree, including depth, node count, and key dependency types. These metrics were selected from the linguistic complexity heuristics from the Universal Dependency set [10].

Overall, these four dimensions are expected to provide complementary perspectives on CQ complexity. A CQ might be semantically complex (e.g., requiring navigation of intricate partonomy or causality relations) yet linguistically simple (e.g., *"What caused this event?"*), scoring high on requirement metrics but low on linguistic/syntactic ones. Conversely, a CQ might lack ontological complexity but is phrased using complex sentence structures, thereby scoring high on syntactic metrics but low on semantic ones.

5. Semantic Overlap: To analyse the semantic characteristics of CQ sets generated by different approaches, we conducted a study on their embeddings. This utilised Sentence-BERT embeddings from the `all-MiniLM-L6-v2` model [11], which generates vectors $\mathbf{e} \in \mathbb{R}^{384}$ capturing the semantic meaning of each CQ (this method follows that adopted by several other studies [3, 5]). Furthermore, to identify semantically equivalent CQs, a pre-defined similarity threshold ($\tau = 0.75$) was determined empirically. This study quantifies the semantic overlap between pairs of CQ sets (e.g., *Set A* $\leftrightarrow$ *Set B*). For each pair, we denote $N_A = |Set\ A|$ and $N_B = |Set\ B|$ as the number of CQs in each set, respectively, and measure:

- **Centroid cosine similarity.** The cosine similarity between the centroids of Set A and Set B provides a measure of the overall alignment of their central semantic representation. A score that is closer to 1 will indicate that the two sets are, on average, focused on similar concepts.
- **Coverage analysis.** We measured how well one set covers the semantic content of another. This was performed in both directions, i.e. for the coverage of Set A by Set B (Set A $\leftarrow$ Set B) we determine:
  - **Mean Maximum Similarity (MMS).** For each CQ embedding $\mathbf{e}_{A,i}$ in Set A, its maximum cosine similarity to any CQ embedding in Set B, $s_{A_i \to B} = \max_j \cos(\mathbf{e}_{A,i}, \mathbf{e}_{B,j})$, was identified. The mean of these $s_{A_i \to B}$ scores (and the standard deviation) indicates how well each CQ in Set A is semantically represented by its closest counterpart in Set B. A higher mean suggests stronger semantic parallels offered by Set B.

[2]Scores were computed using the `textstat` Python library and interpreted comparatively, given the short, interrogative nature of CQs.

Table 1
Pairwise semantic comparison of CQ sets over embeddings (similarity threshold $\tau = 0.75$). Each row compares two sets (*Set A* ↔ *Set B*) based on: centroid cosine similarity (overall thematic alignment), directional coverage (% of Set1's CQs represented by Set2 and vice versa), and MMS (mean of maximum similarities per CQ). Bidirectional coverage measures the overall shared semantic space among sets.

Comparison	Centroid	Set1 ← Set2		Set1 → Set2		BiDirect.
(Set1 ↔ Set2)	Sim.	Cov.(%)	MMS	Cov.(%)	MMS	Cov.(%)
HA-1, HA-2	0.82	**20.5**	0.62 ± 0.15	11.1	0.58 ± 0.15	**15.3**
HA-1, Pattern	0.83	15.9	0.61 ± 0.15	**13.2**	0.57 ± 0.16	14.6
HA-1, GPT	0.85	9.1	0.56 ± 0.13	11.5	0.60 ± 0.12	10.0
HA-1, Gemini	0.73	0.0	0.53 ± 0.11	0.0	0.53 ± 0.11	0.0
HA-2, Pattern	0.84	9.3	0.57 ± 0.15	**13.2**	0.59 ± 0.15	10.9
HA-2, GPT	0.74	1.9	0.48 ± 0.11	3.8	0.55 ± 0.10	2.5
HA-2, Gemini	0.61	0.0	0.46 ± 0.13	0.0	0.51 ± 0.12	0.0
Pattern, GPT	0.77	5.3	0.49 ± 0.16	7.7	0.56 ± 0.14	6.2
Pattern, Gemini	0.68	2.6	0.55 ± 0.11	2.4	0.53 ± 0.12	2.5
GPT, Gemini	0.80	3.8	0.61 ± 0.13	2.4	0.57 ± 0.11	2.9

  - **Set coverage and novelty.** The percentage of CQs in Set A for which $s_{A_i \to B} \geq \tau$ (where τ is the pre-defined similarity threshold described above) was calculated. This quantifies the proportion of Set A's semantic content considered adequately "explained" or represented by Set B. Consequently, the percentage novelty represents the proportion of Set A that introduces semantic content not found (or not closely matched via τ) in Set B.

  The same metrics were computed for the coverage of Set B by Set A.
- **Bidirectional coverage:** This symmetric metric quantifies the overall mutual semantic overlap. It was calculated as $\frac{N^{\text{cov}}_{A \to B} + N^{\text{cov}}_{B \to A}}{N_A + N_B}$, where $N^{\text{cov}}_{A \to B}$ is the number of CQs in Set A covered by Set B (i.e., $s_{A_i \to B} \geq \tau$), and $N^{\text{cov}}_{B \to A}$ is the number of CQs in Set B covered by Set A. Hence, a higher percentage indicates greater shared conceptual space between the two sets.

Together, these dimensions provide a comprehensive, multidimensional view of the suitability, expressiveness, and diversity of CQs produced by different elicitation methods, grounded in both expert assessment and computational analysis.

3. Overview of Evaluation Outcomes

The results for the expert evaluation clearly favoured manually authored CQs. The manual method (HA-1 and HA-2) achieved a mean suitability score of 2.65, with 94.5% of questions accepted by a majority of annotators. This indicates that domain experts are highly effective at producing suitable CQs. The LLM-based methods (GPT-4.1 and Gemini) achieved an average score of 1.24 with 76.0% acceptance, suggesting moderate reliability. Pattern-based CQs scored lowest, with a mean suitability of 0.11 and only 50% acceptance. For readability, human-authored CQs had the lowest FKGL and DCR scores, indicating clearer phrasing. GPT-4.1 generated the most complex and least readable CQs (FKGL 11.64). LLM-generated CQs were also significantly longer (c0), richer in ontological references (c1), and more syntactically complex (c3) than manual or pattern-based ones. Figure 1 consolidates our findings, showing distinct feature profiles from the CQs generated by each elicitation approach.[3]

The pairwise comparison results (Table 1) show that the cosine similarities between the centroids of most pairs are relatively high (typically ranging from 0.61 to 0.85). This suggests that, at a high level, all sets tend to address the same core thematic area defined by the user story. The lowest centroid similarities were observed in comparisons involving `Gemini` (e.g., 0.61 with `HA-2`), indicating its central theme might be slightly more distinct than the other sets.

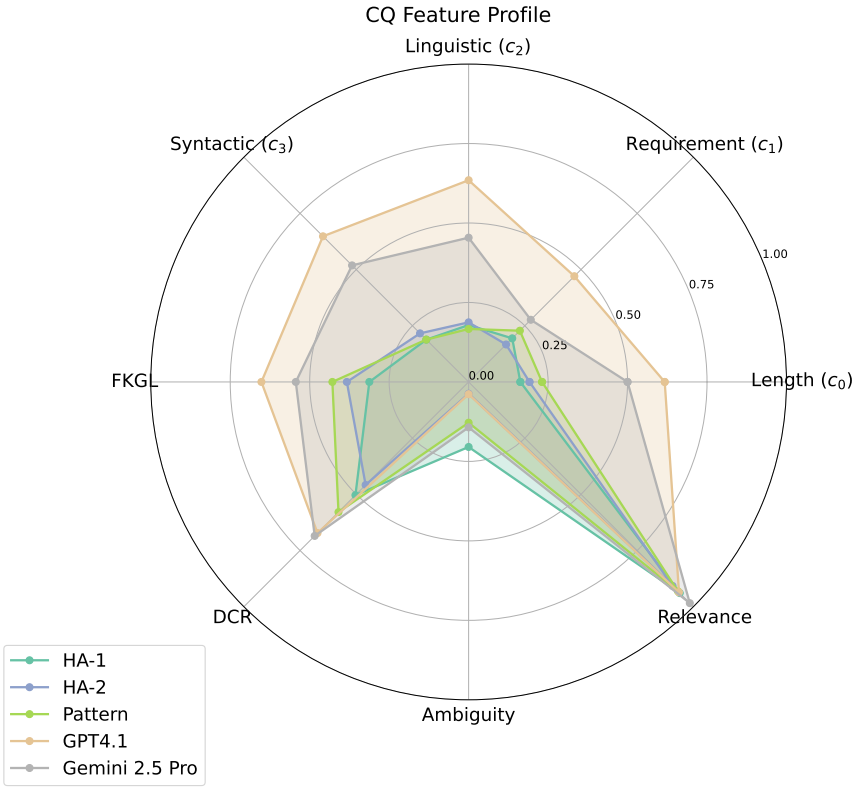

Figure 1: Min-max normalised CQ feature profile per elicitation approach/set in AskCQ.

Despite these relatively high centroid similarities, the specific semantic coverage between sets is low, denoting high degrees of novelty, i.e. a high number of CQs not previously generated. The percentage of CQs in one set covered by another (i.e., having a CQ in the other set with similarity $\geq$ 0.75) is consistently below 21%, and often below 10%. Between the two human annotators ($HA-1 \leftrightarrow HA-2$), who shared a high centroid similarity (0.82), HA-2 covered 20.5% of HA-1's CQs, and HA-1 covered 11.1% of HA-2's CQs (HA-2 has 10 more CQs than HA-1), yielding a bidirectional coverage proportion of 15.3%.

4. Discussion and Conclusion

Our findings suggest that CQs manually crafted by ontology engineers tend to demonstrate the highest suitability for OE, due to achieving better readability, lower complexity, and uniquely capturing inferential requirements (implicit functional requirements) essential for robust ontology design. While LLMs can produce relevant and thematically coherent outputs, the resulting CQs exhibited higher complexity, lower readability, and their semantic coverage, though broad, exhibits limited overlap with human-generated CQs and amongst each other. These results suggest that while LLMs can provide reasonable CQs, these are not comparable to expert authored ones, and that human expertise still remains critical in Ontology Engineering. Crucially, our insights on CQ characteristics and limitations of current automated approaches can be leveraged to directly inform and improve their elicitation methods, aiming to better align their outputs with the desiderata of ontology engineers.

Declaration on Generative AI

Generative AI was only used in the experiments described in the paper, and no Gen AI tool was used to compose or edit the text.

[3]A full discussion of the results and analysis for this evaluation is available in [12].

References

[1] G. K. Q. Monfardini, J. S. Salamon, M. P. Barcellos, Use of competency questions in ontology engineering: A survey, in: Proc. of the Conceptual Modeling: 42nd International Conference, ER, Springer-Verlag, 2023, p. 45–64.

[2] C. M. Keet, Z. C. Khan, Discerning and characterising types of competency questions for ontologies, 2024. URL: https://arxiv.org/abs/2412.13688. arXiv:2412.13688.

[3] R. Alharbi, V. Tamma, F. Grasso, T. R. Payne, A review and comparison of competency question engineering approaches, in: Proc. 24th International Conference on Knowledge Engineering and Knowledge Management, EKAW, Springer Nature, 2024, pp. 271–290.

[4] Y. Ren, A. Parvizi, C. Mellish, J. Z. Pan, K. van Deemter, R. Stevens, Towards competency question-driven ontology authoring, in: Proc. of the 11th Extended Semantic Web Conference, ESWC, Springer International Publishing, 2014, pp. 752–767.

[5] F. Ciroku, J. de Berardinis, J. Kim, A. Meroño-Peñuela, V. Presutti, E. Simperl, Revont: Reverse engineering of competency questions from knowledge graphs via language models, Journal of Web Semantics 82 (2024) 100822.

[6] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, Ontochat: A framework for conversational ontology engineering using language models, in: Proc. of the 21st Extended Semantic Web conference, ESWC, Springer Nature Switzerland, 2025, pp. 102–121.

[7] J. de Berardinis, V. A. Carriero, N. Jain, N. Lazzari, A. Meroño-Peñuela, A. Poltronieri, V. Presutti, The polifonia ontology network: Building a semantic backbone for musical heritage, in: Proc. of the 22nd International Semantic Web Conference, ISWC, Springer, 2023, pp. 302–322.

[8] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, B. S. Chissom, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Technical Report Research Branch Report 8-75, Naval Air Station Memphis, Research Branch, Millington TN, 1975.

[9] E. Dale, J. S. Chall, A Formula for Predicting Readability: Instructions, Ohio State University Bureau of Educational Research, 1948.

[10] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, C. D. Manning, Universal Stanford dependencies: A cross-linguistic typology, in: Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 4585–4592.

[11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[12] R. Alharbi, V. Tamma, T. R. Payne, J. de Berardinis, A comparative study of competency question elicitation methods from ontology requirements, 2025. URL: https://arxiv.org/abs/2507.02989. arXiv:2507.02989.