

Agentic Competency Question Generation for Copolymer Domain Ontologies

Vishvapalsinhji Parmar^{1,*}, Vamsi Krishna Kommineni^{1,†} and Birgitta König-Ries¹

¹Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Jena, Germany

Abstract

Competency questions (CQs) play a central role in ontology engineering, yet their authoring remains a significant bottleneck. Their creation typically requires the involvement of several domain experts, resulting in a very time-consuming, laborious process. An alternative is to extract them from a representative set of papers. This, however, is equally labor-intensive when done manually. Thus, scaling CQ creation across a domain literature corpus demands automation. We present an agentic retrieval-augmented pipeline that automatically generates, validates, and consolidates domain-level CQs from scientific papers. The pipeline operates in two phases. In Phase A, each paper is distilled into a bounded, evidence-bearing knowledge state (a rolling structured summary) through iterative extraction, and persona-conditioned CQs are generated from retrieved evidence passages across seven domain personas, ensuring that every question is grounded in the source text rather than in model priors. In Phase B, the per-paper sets are progressively narrowed through deduplication, cross-paper generalization, RAG-based validation, and two-stage consolidation into ontology-ready CQs carrying triple patterns, typed variables, and SPARQL sketches. We tested this pipeline on a corpus of 14 expert-curated papers on supercritical-CO₂ polymer foaming, the pipeline produced 682 CQs with a mean question faithfulness of 0.98 and zero off-paper questions, consolidating them into 120 domain-level CQs, an 82.4% reduction.

Keywords

CQ Generation, Agentic Pipeline, Ontology Engineering

1. Introduction and Motivation

Competency Questions are natural-language questions that an ontology is expected to answer. They have been used as a requirements artifact in ontology engineering since Grüninger and Fox proposed them for enterprise-integration ontologies [1]. A recent survey of ontology engineers found that CQs are used mainly to define an ontology's scope and to evaluate its conceptualization [2]. The difficulty is that producing a good set of CQs, i.e., one that covers all relevant aspects of the domain, by hand takes real effort. The CORAL corpus, for example, gathered 834 functional requirements across 14 ontologies [3].

To reduce this effort, several groups have proposed generating CQs with large language models (LLMs) from domain documents. For example, D. Di Nuzzo et al. (2024) [4] show that LLMs can produce reasonable CQs at scale from technical standards. Applying this to scientific papers, however, raises further limitations. The first is that providing a full paper in a single prompt is unreliable, even with long-context models. Liu et al. [5] report a U-shaped pattern in which information placed in the middle of a long context is used much less reliably than information near the start or end, so a long paper does not give uniform grounding. We therefore generate CQs using retrieved passages rather than from the paper as a whole. The second is persona collapse, where a single expert prompt yields homogeneous questions. In reality, a domain knowledge graph is used by different people who bring different questions to it. In our polymer use case, synthesis chemists, material characterization scientists, ontology engineers, and students all have distinct information needs. A single generic prompt does not capture that range, so we condition generation on these reader personas. The third is sycophancy,

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

*Corresponding author.

[†]These authors contributed equally.

✉ vishvapalsinhji.parmar@uni-jena.de (V. Parmar); vamsi.krishna.kommineni@uni-jena.de (V.K. Kommineni); birgitta.koenig-ries@uni-jena.de (B. König-Ries)

ORCID 0000-0002-4370-2729 (V. Parmar); 0000-0001-6168-3085 (V.K. Kommineni); 0000-0002-2382-9722 (B. König-Ries)



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

which occurs when an LLM prioritizes agreeing with a user’s stated beliefs over providing truthful or objective information. For instance, Sharma et al. demonstrate that several assistants will readily agree with a user’s stated view across a variety of tasks [6]. When validation and refinement are also performed by an LLM, related issues emerge, such as a preference for the model’s own style of answer [7] alongside biases of position and verbosity [8]. Without proper oversight, a judge who merely agrees with the prompt and a refiner who unquestionably rewrites on request will quietly corrupt the generated question set. Most existing works on LLM-generated competency questions assess quality by comparing generated questions against a gold standard CQ set. However, in our setting, no such reference set is available for the copolymer domain targeted by our pipeline. The Validating Semantic Pitfalls in Ontology (VSPO) paper highlights that evaluation based purely on similarity to benchmark datasets risks missing the modeling defects that are most relevant in practice [9]. Likewise, Alharbi et al. point out that cross-study comparison is challenging due to variation in tasks, datasets, and evaluation metrics, as well as the frequent absence of the competency questions used during ontology development [10].

These limitations suggest that evaluation should focus less on surface similarity to reference questions and more on whether the generated questions expose meaningful modeling requirements. To address these challenges, this paper introduces a novel pipeline for systematically generating and evaluating competency questions without relying on a pre-existing gold standard. The core contributions of this work are: **Iterative Grounded Extraction:** an extraction approach that processes the literature sequentially to build a running structured summary. To ensure traceability, every extracted item is explicitly tied to a verbatim source sentence, its surrounding context, and a confidence band. **Decoupling Persona from Question Structure:** our generation methodology crosses seven distinct reader personas, each mapped to a cognitive level from the revised Bloom taxonomy [11] with ten knowledge-graph-aligned question archetypes. Here, a persona defines who is asking, while an archetype defines the structural shape of the question. For example, the Student persona asks at a beginner’s recall level, while the ENTITY_LOOKUP archetype shapes the question as ”What is the property of material?”, combining the two gives a simple question like ”What is the glass transition temperature of this polymer?”, and the same archetype can instead be paired with the Synthesis Chemist persona so that any persona can be combined with multiple question forms rather than being locked to a single template. This process is grounded via retrieval over a local FAISS index [12] utilizing BGE-small embeddings [13]. **Cross-Model Validation and Refinement:** the validation step mitigates sycophancy by employing different model families at each pipeline stage (Qwen for extraction, Gemma for generation, DeepSeek for evaluation, and Qwen for refinement). The prompt design explicitly permits the judge to approve outputs without edits and allows the refiner to reject unnecessary changes. **Reference-Free Evaluation Framework:** the four-metric reference-free evaluation framework measures Coverage, Question Faithfulness, Diversity, and Triplifiability and scores the generated question set against the retrieved evidence passages and the structured knowledge state extracted from each paper. This serves as our direct substitute for the missing-gold-set problem highlighted in prior work [9, 10]. **Corpus-Level Consolidation:** the deduplication and consolidation step aggregates 14 independent, paper-specific question sets into a unified, schema-level catalog. This consolidated catalog is designed to directly support the development of a domain-specific polymer ontology.

2. Agentic CQ Generation Pipeline

Generating competency questions manually requires a significant amount of effort from human experts, involving back-and-forth interaction between domain experts (i.e., chemists) and ontology engineers. Our pipeline focuses on leveraging the capabilities of Large Language Models (LLMs) with an agentic approach. We frame automatic CQ generation as a two-phase (Phase A and Phase B) process operating over a domain corpus shown in Figure 1. The core design is organized around two principles: every CQ is grounded in retrieved textual evidence rather than the model’s parametric memory, and CQs are progressively refined from per-paper sets into a consolidated, domain-level set with full provenance. Whereas previous approaches derived CQs from existing ontologies [14], we use domain literature as

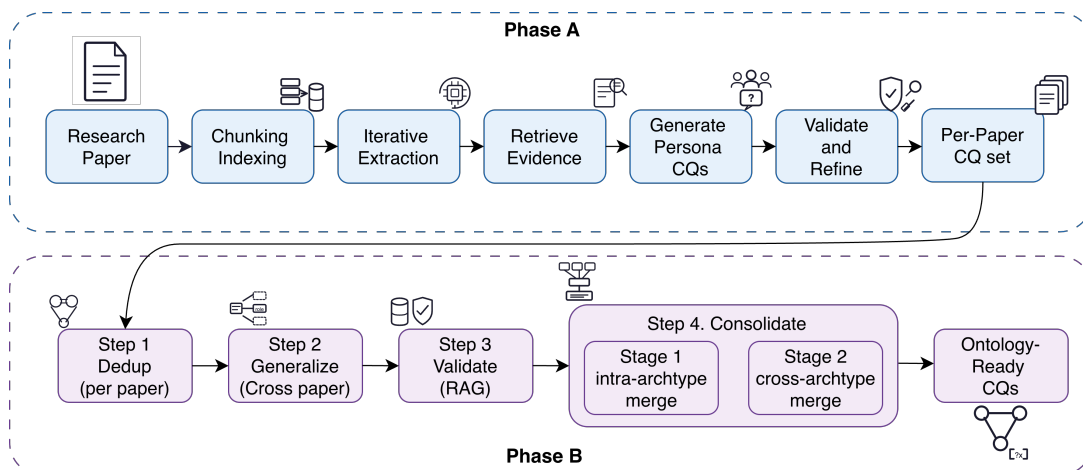


Figure 1: Two-phase agentic pipeline for competency question generation. **Phase A** generates persona-conditioned CQs from each paper using retrieval-augmented evidence. **Phase B** narrows and consolidates them across the corpus into ontology-ready domain-level CQs.

input, incorporating a retrieval-augmented setup with multi-agent decomposition, persona conditioning, validation and refinement, and cross-paper consolidation.

2.1. Phase A - Per Paper Generation

Phase A assumes the existence of a corpus of papers covering the domain of interest. For now, we rely on experts to provide this corpus; alternatively, emerging tools like the ASTA tool suite (<https://allenai.org/asta/>) could be used. Each paper is first cleaned and segmented using a recursive character splitter into overlapping chunks. These chunks are then embedded and stored in a per-paper FAISS index. The core of this phase is an iterative knowledge-extraction agent that traverses the chunks while maintaining a bounded rolling state. For each chunk, the prompt combines static domain priors, the compacted state accumulated so far, and the current chunk. The agent then emits structured items in four categories: concepts, relations, properties, and findings. Each category entry includes an evidence sentence, surrounding context, and a calibrated confidence score. The extracted items are deduplicated and merged into the state under per-category caps, ensuring that an arbitrarily long paper is distilled into a compact, evidence-based knowledge representation without requiring the full text in a single prompt. A retrieval agent then selects the most important concepts and retrieves passages relevant to them. Following this, Competency Question (CQ) generation is performed once per persona. For each of seven domain personas (e.g., synthesis chemist, or material characterization scientist), a generation agent receives four inputs: the persona description, the structured knowledge state, a set of passages retrieved with that persona’s interests in mind, and a small library of question templates called archetypes. The agent then generates a set of questions. Each generated question is annotated with its archetype, expected answer, persona code, and source section from the paper. Additionally, a coverage-gap pass is performed to issue one targeted round per persona over any extracted concepts not addressed in the initial generation. Following this, the generated CQs are validated by a validation agent combining rule-based checks (such as persona coverage and concept usage), embedding-based near-duplicate detection, and a per-question LLM-based judge. Each question is scored on answerability and specificity against semantically selected evidence. Questions that fail validation are passed to a refinement agent, which, guided by the judge’s feedback and relevant evidence, either revises, retains, or flags them as unrecoverable. Finally, the refined set of CQs is produced for each paper.

2.2. Phase B - Cross-paper narrowing

Once each paper has its own set of competency questions, those sets are combined across the corpus in four steps. The first step is **deduplication**, which is applied once per paper. A deduplication agent merges questions that ask essentially the same thing, but only within the same archetype. Every removed question is logged with a pointer to the question that absorbed it, so the edit trail stays

auditable. The second step, **generalization**, compiles once per archetype across the whole corpus. A generalization agent groups questions that ask the same kind of thing across papers and replaces paper-specific entities with their domain role. For example, a specific polymer (e.g., PMMA) becomes simply "amorphous polymer." The output is a set of role-level questions instead of entity-level ones. The third step is **validation by retrieval**, each generalized question is checked against the FAISS indices built in Phase A. In a single call, the validation agent fetches the most relevant passages, generates a candidate answer, judges its fidelity to the evidence, and assigns a quality flag (strong, partial, over-generalized, or paper-specific). The fourth step is **consolidation**, run in two stages. Stage 1 merges questions within each archetype by their underlying triple signature, same (subject-class, predicate, object-class) shape, replacing differing fillers with typed variables. Stage 2 then merges across archetypes wherever the same triple signature still appears, producing the final set of domain-level questions, each carrying a triple pattern, typed variables, a SPARQL sketch, and the schema fragments it implies. Every consolidated question records the identifiers of all source questions it absorbed, and a semantic audit flags any unaccounted or hallucinated identifiers and any residual duplicates.

The pipeline assigns distinct, single-responsibility roles, i.e., extractor, retriever, generator, judge, refiner, consolidator, across four open model families (Qwen, Gemma, DeepSeek, Llama), with schema-validated hand-offs between stages. Using deliberately different model families for generation and judging reduces self-preference effects in LLM-based evaluation [8]. The pipeline runs entirely on open models served via the university’s high-performance computing infrastructure¹, making the full process reproducible without dependence on proprietary APIs. To further support reproducibility, the full pipeline code, configuration files, and evaluation scripts are publicly available².

3. Evaluation

We applied the pipeline to a corpus of 14 papers on supercritical-CO₂ foaming of amorphous polymers and blends, curated by Marcel Sticher (Universität Bayreuth) within the COIN project³. Because the deliverable is a set of questions, not their answers, we evaluate quality with a reference-free protocol that scores four families, i.e., coverage of the extracted knowledge state, question faithfulness (semantic grounding of each question in retrieved evidence), diversity, and triplifiability, and report two diagnostics (answer verifiability and specificity) that are computed but not scored.

Per-paper CQ quality. Across the corpus of 682 generated CQs, mean question faithfulness was 0.98, with *zero* off-paper questions in total: every generated question was judged answerable from the source paper’s own evidence (Table 1). This is direct evidence that grounding generation in retrieved evidence, rather than in model priors, suppresses hallucinated questions. Coverage and diversity⁴ are the current limitations, which are targeted by future work.

Metric	Mean	SD	Range
Question faithfulness	0.98	0.01	0.96–1.00
Coverage	0.48	0.08	0.32–0.65
Diversity	0.29	0.02	0.25–0.32
Triplifiability	0.68	0.04	0.63–0.76
Overall score	0.69	0.03	0.62–0.76

Table 1

Reference-free validation results of the pipeline (n = 682 CQs, 14 papers). Mean question-to-evidence cosine is 0.81. SD is Standard Deviation.

Cross-paper narrowing. From the 682 generated CQs, consolidation produced 536 deduplicated per-paper CQs, 490 generalised CQs, and finally 120 domain-level CQs, an 82.4% reduction across 11 LLM

¹Models served via Ollama on the Draco HPC endpoint at the FSU Jena at ollama.draco.uni-jena.de

²<https://github.com/fusion-jena/Agentic-CQ-Generation>

³COIN Project, Group C1: <https://www.coin.uni-jena.de/en/253/project-c1>

⁴Evaluation metrics: <https://github.com/fusion-jena/Agentic-CQ-Generation#evaluation-metrics>

calls. The consolidated CQs are ontology-ready: a single causal CQ, “How does $\langle entity \rangle$ affect $\langle property \rangle$?”, subsumes 34 source questions spanning 12 of the 14 papers, capturing a recurring affects(entity, property) pattern as one reusable, schema-bearing question.

4. Conclusion & Future work

We have presented an agentic, retrieval-augmented pipeline that turns a domain literature corpus into consolidated, ontology-ready competency questions. By grounding every question in retrieved evidence and narrowing per-paper questions into a provenance-preserving domain set, the pipeline produces highly faithful CQs (0.98 mean faithfulness, no off-paper questions across 682 CQs) while compressing them by over 80% into 120 reusable domain-level questions.

Several directions follow naturally. First, the consolidated triple patterns and SPARQL sketches are intended as scaffolding for ontology construction; driving schema generation directly from them would close the loop from literature to ontology. Second, while our reference-free protocol avoids dependence on a gold standard, validating the generated CQs against expert-authored questions would complement it.

Acknowledgments

VP’s work is funded by DFG within GRK 3040 (COIN) (project number 527537972).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to perform grammar and spelling checks, improve writing style, and generate code snippets. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication’s content.

References

- [1] M. Grüninger and M. S. Fox. “The Role of Competency Questions in Enterprise Engineering”. In: *Benchmarking - Theory and Practice*. Ed. by A. Rolstadås. Boston, MA: Springer US, 1995, pp. 22–31.
- [2] G. K. Q. Monfardini et al. “Use of competency questions in ontology engineering: A survey”. In: *International Conference on Conceptual Modeling*. Springer. 2023, pp. 45–64.
- [3] A. Fernández-Izquierdo et al. “CORAL: A Corpus of Ontological Requirements Annotated with Lexico-Syntactic Patterns”. In: *Extended Semantic Web Conference*. 2019.
- [4] D. Di Nuzzo et al. “Automated Generation of Competency Questions Using Large Language Models and Knowledge Graphs”. In: *NLP4KGC@SEMANTiCS*. 2024.
- [5] N. F. Liu et al. “Lost in the Middle: How Language Models Use Long Contexts”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173.
- [6] M. Sharma et al. “Towards Understanding Sycophancy in Language Models”. In: *International Conference on Learning Representations*. Ed. by B. Kim et al. Vol. 2024. 2024, pp. 110–144.
- [7] A. Panickssery et al. “LLM evaluators recognize and favor their own generations”. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. 2024.
- [8] L. Zheng et al. “Judging LLM-as-a-judge with MT-bench and Chatbot Arena”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023.
- [9] H. Choi et al. “VSPO: Validating Semantic Pitfalls in Ontology via LLM-Based CQ Generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. 23. 2026, pp. 18977–18984.
- [10] R. Alharbi et al. “Characteristics and Desiderata for Competency Question Benchmarks”. In: *HGAIS@ISWC*. 2024.
- [11] C. Gonsalves. “Generative AI’s Impact on Critical Thinking: Revisiting Bloom’s Taxonomy”. In: *Journal of Marketing Education* 48 (2024), pp. 4–19.
- [12] M. Douze et al. “The Faiss Library”. In: *IEEE Transactions on Big Data* 12.2 (2026), pp. 346–361.
- [13] S. Xiao et al. “C-Pack: Packed Resources For General Chinese Embeddings”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024, pp. 641–649.
- [14] Y. Rebboud et al. “Can LLMs Generate Competency Questions?” In: *ESWC Satellite Events*. 2024.