Mind the Gap: LLMs, Competency Questions, and the Non-Technical User in the Humanities Domain

Claire McNamara[1,*], Lucy Hederman[1] and Declan O'Sullivan[2]

[1]Trinity College Dublin, Ireland
[2]ADAPT Centre for Digital Content, Trinity College Dublin, Ireland

Abstract

Non-technical users often face a significant barrier when first attempting to explore complex knowledge graphs (KGs). We define this challenge as the Initial Exploration Problem, characterised by three interrelated barriers: ontology opacity, query incapacity, and scope uncertainty. This paper investigates how large language models (LLMs) can support domain experts in addressing this problem by automatically generating template-style competency questions (CQs) for the Virtual Record Treasury of Ireland (VRTI) KG. These templates are not user-facing themselves, but serve as scaffolding for creating curated questions (CuQs), expert-validated, natural language questions that help new users begin meaningfully exploring the graph. We evaluate two LLMs (GPT-4o and Gemini 2.0 Flash) across twelve prompt configurations varying in scope and framing, and assess question quality using both semantic similarity to expert-authored CQs and detailed expert review. Our findings highlight how prompt design influences LLM output, and underscore the value of combining automated generation with expert curation. Ultimately, we propose a practical pipeline to support the creation of exploratory entry points tailored to user needs, helping domain experts craft better questions, and helping users take their first steps into meaningful KG exploration.

Keywords

Knowledge Graph Exploration, Competency Questions, Large Language Models, Non-Technical Users, Initial Exploration Problem

1. Introduction

A Knowledge Graph (KG), as defined by Hogan et al. in 2021 [1], is "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities". KG technologies have generated a diverse range of research topics over the last two decades, from event-based networking [2], to more recently their security [3]. However, by far the most common deployment has been for data integration, as it is a powerful way of representing the connections between information across diverse datasets, such as those found in the humanities, that would ordinarily remain separate. It affords those interested in the humanities the potential to identify patterns in data that may not be otherwise obvious when exploring such datasets through more traditional forms (e.g., CSV, NoSQL) [1]. Utilising the W3C Resource Description Framework (RDF) model [4], it is possible to represent the links between data across the datasets in a machine-readable format (RDF triples), thereby functionally combining the KG datasets in a meaningful way. The associated query language, SPARQL [5], can then be used over the RDF to return answers to questions users may have about the information in the dataset.

However, a significant challenge arises when non-technical people attempt to explore an unfamiliar humanities KG [6]. This challenge, which we define as the ***Initial Exploration Problem***, describes the difficulty faced by non-technical individuals who do not know where to begin when confronted with a large, complex, and unfamiliar KG. To our knowledge, while this problem has been addressed implicitly in recent work [7], it has not yet been formally defined in the literature.

First International Workshop on Users and Knowledge Graphs (UKG), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

[*]Corresponding author.

mcnamacl@tcd.ie (C. McNamara); hederman@tcd.ie (L. Hederman); declan.osullivan@adaptcentre.ie (D. O'Sullivan)
0000-0002-8263-8694 (C. McNamara); 0000-0001-6073-4063 (L. Hederman); 0000-0003-1090-3548 (D. O'Sullivan)

The Initial Exploration Problem is characterised by three interrelated difficulties:

1. **Ontology Opacity**: Inability to interpret class/property terms and hierarchies (e.g., abstract event types such as `E13_Attribute_Assignment` or `E65_Creation` from the CIDOC CRM ontology [8], a widely used ontology standard in cultural heritage domains).
2. **Query Incapacity**: Lack of skills to formulate structured queries using formal query languages such as SPARQL [5].
3. **Scope Uncertainty**: No prior awareness of what questions the KG can answer. This is not simply a matter of vague search intent or query refinement, as in traditional exploratory search [9], nor is it addressed by standard usability solutions. Rather, it is a problem of orientation: the user may not know what kinds of questions are answerable, how the data is structured, or where to begin their exploration.

This gap presents a substantial barrier to new users engaging with KGs like the Virtual Record Treasury of Ireland (VRTI) [10] KG [11], which is the focus of our experiment. The VRTI KG provides a strong use case for this experiment as it currently contains over 2.7 million triples and is constructed using complex ontologies such as CIDOC CRM [8], GeoSPARQL [12], and its own domain-specific VRTI ontology [13]. Previous studies have shown that without a clear understanding of the information stored in the KG [14] or the technical skills to create and execute SPARQL queries [15][16], those interested in the humanities often struggle to explore the data effectively.

Tús Maith (/tuːs ˈmˠah/), coming from the Irish saying "Tús Maith Leath na hOibre" ("a good start is half the work"), is a framework designed to address the Initial Exploration Problem. A key component of Tús Maith is the concept of curated questions (CuQ). In this context, CuQs are a set of predefined natural language questions curated by a domain expert to guide new users in beginning their exploration of the KG. These questions act as a set of exploration starting points for new users while also giving them an understanding of the types of questions the KG is able to answer. As will be explained in Section 2, a past experiment carried out by the authors found that CuQs show promise when used in this context. These CuQs are conceptually linked to competency questions (CQs), which are traditionally used in ontology engineering to define the requirements a knowledge base should be able to meet [17]. While CQs are typically designed to guide ontology development or evaluate knowledge completeness [17], in this work, they are adapted as a mechanism for supporting user exploration.

This paper presents an experiment on the automatic generation of template-based CQs from the VRTI KG using two popular Large Language Models (LLMs): ChatGPT-4o [18] and Gemini 2.0 Flash [19]. These templates are not intended as final CuQs, but as an intermediate resource: structured natural language forms grounded in the VRTI KG's classes, properties, and, where they exist, relationships. They are designed to be later filled with specific, commonly searched entities to produce fully-formed questions suitable for non-technical users in an exploration interface. The aim is not to develop new prompt engineering strategies or fine-tuning methods. Rather, its contribution lies in combining previously separate research threads (LLM-based CQ generation and KG exploration techniques) and applying them in a holistic way. Specifically, it explores how well a by-product of CQs can be repurposed as a solution to the Initial Exploration Problem.

2. Tús Maith

As seen in Figure 1, the template-style CQs are intended to support domain experts in crafting CuQs that help users explore the scope and potential of the KG. The lifecycle of a question in this process involves three stages:

1. **Template-style competency question (CQ)**: A general template question generated by the LLM based on class/property input, e.g. *"Where was a person born?"*.
2. **Filled-in/generated question**: A template-style CQ filled in automatically with popular entities within the graph and/or entities the domain expert feels may be of interest, e.g. *"Where was Oscar Wilde born?"*.

3. **Curated question (CuQ)**: The filled-in question accepted or refined by a domain expert for presentation to users.

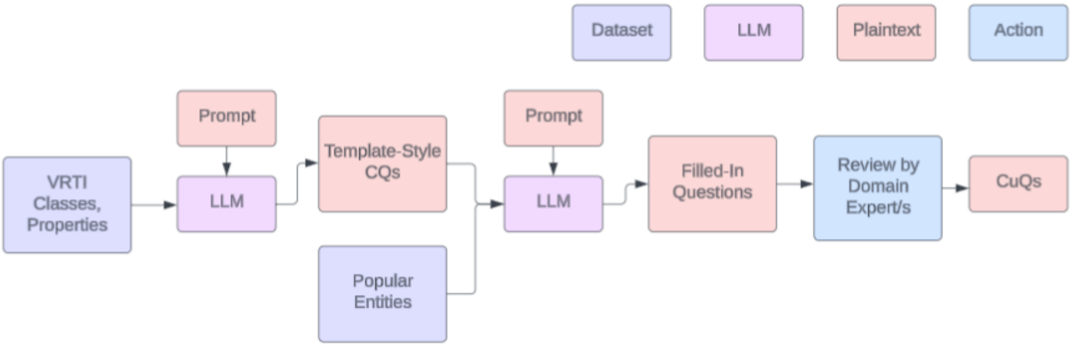
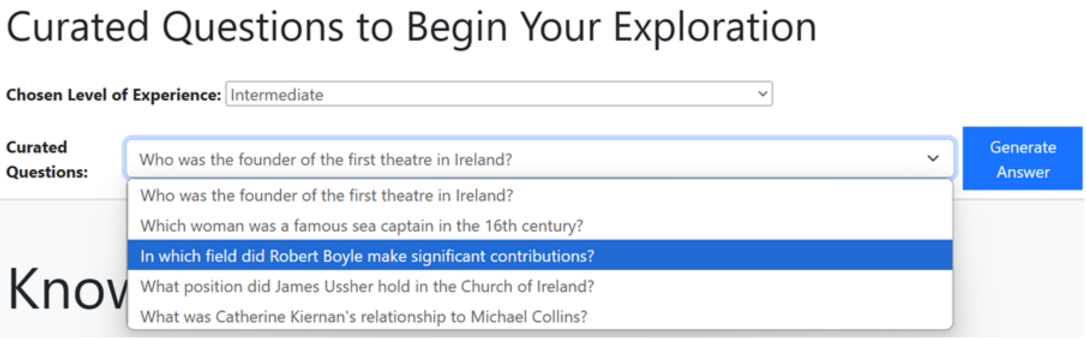


Figure 1. The Tús Maith pipeline: from LLM-generated templates to curated user-facing questions.

Prior to the experiment outlined in this paper, a small-scale evaluation involving 12 participants was conducted by the authors to assess the current user-facing components of Tús Maith. These components, which were implemented within the VRTI Explorer interface [20], were CuQs and commonly searched keywords, used in conjunction with a search bar. An example of how CuQs were presented can be seen in Figure 2. The CuQs for this evaluation were created by a historical domain expert independent of Tús Maith, who has been heavily involved in the development of the VRTI KG. This expert was asked to formulate fifteen questions that they considered both interesting and illustrative of the VRTI KG's scope [21]. The results of the evaluation showed that three of the four participants who self-identified as having limited familiarity with the VRTI KG found the CuQs to be the most helpful feature when beginning their exploration. While a further, larger-scale evaluation targeting more novice users is needed to better assess the efficacy of CuQs in addressing the Initial Exploration Problem, the preliminary findings were promising and motivated further investigation into scalable methods of generating such questions.


Curated Questions to Begin Your Exploration

Chosen Level of Experience: Intermediate

Curated Questions: Who was the founder of the first theatre in Ireland?

Who was the founder of the first theatre in Ireland?
Which woman was a famous sea captain in the 16th century?
In which field did Robert Boyle make significant contributions?
What position did James Ussher hold in the Church of Ireland?
What was Catherine Kiernan's relationship to Michael Collins?

Figure 2. The curated questions integrated into the interface as a dropdown menu.

3. Experiment Methodology

This section outlines the design of the experiment and the two-step evaluation process that was followed in order to explore the extent to which LLMs at default settings (temperature=1, top_p=1) can be used to generate template-style CQs for the VRTI KG. The design was motivated by the methodology presented in [22], which evaluated six LLMs across five RDF-based ontologies using zero- and few-shot prompting. While our experiment involved only two LLMs, a greater variety of prompts was used (12 distinct prompts as opposed to 6), and crucially, domain experts evaluated the output. The full results and code are openly available at [21], [23] ensuring transparency and reproducibility.

3.1. Prompt Design and Experimental Setup

Twelve separate runs were conducted for each LLM, varying by prompt framing, shot type (one-shot or few-shot), and input scope as seen in Table 1. Input scope options are discussed in section 4.2.

1. **Prompt 1**: Simply instructs the LLM to generate questions from provided classes and properties.
2. **Prompt 2**: Includes the assigning of a role to the LLM (historical knowledge graph exploration assistant) to guide its generation strategy.

Table 1
The run numbers corresponding to each combination of prompt type, shot type, and input scope.

Prompt Type, Shot Type	All Classes & Properties	Subset	Subset + CIDOC CRM Relationship Information
Prompt 1, One-shot	Run 1	Run 2	Run 3
Prompt 1, Few-shot	Run 4	Run 5	Run 6
Prompt 2, One-shot	Run 7	Run 8	Run 9
Prompt 2, Few-shot	Run 10	Run 11	Run 12

Listing 1 shows the base prompt used for all LLM runs, with conditional sections annotated according to experimental configurations (e.g., +Role, +Examples).

Listing 1: Template-style competency question prompt with condition annotations.

```
# Template-style competency question prompt.
# Modular components annotated with experimental conditions.
# --------------------------------------------------------

{INCLUDE_ROLE? You are a historical knowledge graph exploration assistant. Your task is to
    generate natural language, template-style competency questions that ordinary users might ask
     when exploring a historical knowledge graph.} [Condition(s): Prompt 2,
    +Role]

I am providing you with a set of knowledge graph classes, properties {INCLUDE_RELATIONSHIPS? ,
    and relationships}. Using only the provided information, generate 20 template-style
    competency questions that could be asked about this ontology. [Condition(s): All;
    relationships only in Subset and CIDOC CRM Relationships, +Relationships]

A competency question in this context is defined as a natural language question that helps
    convey the scope and potential of a KG to users unfamiliar with its structure with the aim
    of supporting exploration.

{INCLUDE_EXAMPLES? Examples of the type of questions I am looking for are: "Where was a person
    born?", "When did a person die?", "What type of place is this?".} [Condition(s): Few Shot,
    +Examples]

Phrase questions in natural, user-friendly language, avoiding technical or ontology-specific
    terms e.g. use "was born" instead of "birth event".

Use template-style phrasing: general questions with placeholders like "a person", "a place" etc.

Focus on creating general, reusable question templates that can later be filled with specific
    instance data.

Only use the given classes, properties {INCLUDE_RELATIONSHIPS? , and relationships}; do not
    invent additional concepts. [Condition(s): All; relationships only in Subset and CIDOC
    CRM Relationships, +Relationships]

You may combine classes and properties if it helps express a useful question.
```

```
Phrase each competency question naturally, as if it were intended for a user exploring the
    knowledge graph.
```

3.2. Data Preparation

The VRTI KG integrates multiple ontologies, including CIDOC CRM [8], GeoSPARQL [12], and a custom VRTI ontology [13], making it significantly more complex than single-schema KGs. CIDOC CRM, in particular, is a formal ontology designed to represent cultural heritage information in rich detail, featuring a dense network of abstract classes and properties that model events, entities, and relationships. The resulting heterogeneity obtained through the use of multiple, often complex ontologies poses additional challenges for natural language question generation, as many terms are either highly abstract or domain-specific (e.g. `P81b_begin_of_the_end`). To explore whether different input scopes could yield more interpretable and relevant questions, three variations of input term sets were tested across prompt runs: the full set of classes and properties (48 classes, 198 properties) [21], a manually reduced subset of 18 classes and 35 properties [21], and a third version of the subset that included CIDOC CRM relationship information to the subset of classes and properties as prompt variables. The subset was constructed by the authors to retain the most semantically expressive and user-relevant terms while excluding overly technical entries, such as `virtrdf#qmf01blankOfShortTmpl`, which are poorly suited to natural language articulation. The decision to extract classes and properties from instance data rather than from the full ontology was intentional; not all ontology terms are currently used in the KG, and this work aims to support exploratory access to available content, rather than exhaustively covering the ontology itself.

3.3. Evaluation Strategy

This subsection outlines the two-step evaluation process undertaken by the authors. The first step involved using SentenceBERT [24] to identify, for each LLM, the set of generated questions most semantically similar to a ground truth set of CQs in order to filter the question sets from 24 to 2. In the second step, those two question sets were analysed by domain experts with intimate knowledge of the VRTI KG, evaluating them across three dimensions: relevance, clarity, and answerability. They were also asked to evaluate whether the generated set of CQs contained questions that, in their opinion, could be added to the ground truth question set.

3.3.1. Step 1 - SentenceBERT Evaluation

There is no existing benchmark CQ set for the VRTI KG. To establish a minimal ground truth, a historical domain expert with direct involvement in the KG's development was asked to create a set of CQs [21]. Generated template-style CQs were first evaluated for semantic similarity to the expert-authored ones with SentenceBERT using the all-mpnet-base-v2 sentence-transformer model [25] and compared pairwise using cosine similarity.

Let $S \in \mathbb{R}^{m \times n}$ represent the similarity matrix, where m is the number of generated questions and n the number of expert-authored ones. Each element s_{ij} corresponds to the similarity between the i-th generated question and the j-th expert question.

For each generated question, the maximum similarity to any expert question was calculated:

$$\mathrm{max}_i = \max_{1 \leq j \leq n} s_{ij}$$

The average of these maximum similarity scores across all generated questions was then computed as:

$$\overline{\mathrm{max}} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{max}_i$$

To assess how many generated questions achieved a similarity above a chosen threshold τ, we calculate:

$$c = \sum_{i=1}^{m} \mathbb{I}(\max_i \geq \tau)$$

where $\mathbb{I}$ is the indicator function. The percentage of generated questions exceeding the threshold is given by:

$$p = \left(\frac{c}{m}\right) \times 100$$

In this experiment, a threshold of $\tau = 0.6$ was selected, following precedent established in prior work evaluating LLM-generated CQs using SentenceBERT[22]. This value reflects a moderate level of semantic alignment, capturing cases where generated questions are meaningfully similar to expert-authored ones while avoiding overly lenient matches.

3.3.2. Step 2 - Expert Evaluation

While semantic similarity provides an automated means of filtering and ranking generated questions, it does not assess important qualitative aspects such as interpretability, alignment with the KG's intended usage, or answerability using only the information found within the KG. To address this, a human expert evaluation was conducted on a subset of the generated questions. Two historical domain experts, both very familiar with the content and structure of the VRTI KG, were asked to evaluate a sample of the LLM-generated questions. This sample comprised of one set of template-style CQs per LLM; the sets of questions that achieved the highest average semantic similarity to the expert-authored ground truth questions for each LLM out of the 12 prompt configurations described in Section 4.1. Each question was evaluated independently by both experts using a 5-point Likert scale [26] across three dimensions:

1. **Relevance**: How well does the question align with the purpose of a CQ in this context? (1 = Not relevant at all, 5 = Highly relevant)
2. **Clarity**: Is the question phrased in a clear, unambiguous, and grammatically correct manner? (1 = Very unclear or confusing, 5 = Very clear)
3. **Answerability**: Can the question be reasonably answered using information contained within the VRTI KG? (1 = Not answerable at all, 5 = Fully answerable)

These three dimensions were selected to reflect the practical requirements of template-style CQs within the context of the Tús Maith pipeline; question templates that highlight the scope and potential of a KG that can later be filled in with specific entity information. *Relevance* considers whether the question aligns well with the definition of a template-style CQ. *Clarity* addresses interpretability, particularly for non-technical users. *Answerability* considers whether the question can reasonably be answered using only the information currently present in the KG.

The experts were instructed to evaluate each question as a template, not a fully instantiated question. In addition to scoring individual questions, the experts were also invited to reflect on whether any questions present in the generated set but not in the original ground truth set could, in hindsight, be considered template-style CQ candidates. This qualitative feedback was intended to identify whether the LLM could highlight potential blind spots in the expert-authored ground truth CQs. Expert scores were aggregated to provide average scores per dimension across each LLM-generated set. This enabled comparison not only between individual questions but also between the two different LLMs used. This evaluation stage was essential in order to identify whether either/both/or neither LLM could produce not only semantically aligned questions with the ground-truth CQs, but ones that were perceived by experts as meaningful, understandable, and practically useful in the context of supporting exploration within a complex KG.

4. Results

4.1. SentenceBERT Similarity Evaluation

We first evaluated the similarity between 20 questions generated in each run and a set of 10 domain expert-created ground truth questions using SentenceBERT. The metric used was maximum similarity per question, with average values reported per run (Table 2). Across all runs, the highest-performing runs were Run 6 (the few-shot prompt with a subset of the classes, properties and CIDOC CRM relationship information) for GPT-4o, and Run 11 (the few-shot prompt with a role assigned and just the subset of classes and properties) for Gemini 2.0 Flash. The best run for GPT achieved an average max similarity of 0.418 (20% ≥ 0.6), while the best for Gemini achieved 0.480 (25% ≥ 0.6), indicating a closer lexical and semantic match to the expert-authored questions.

Table 2
Average semantic similarity score (as computed using SentenceBERT) per run, and the number of generated questions per run that met or exceeded the 0.6 similarity threshold.

Run	GPT Avg. Sim	GPT ≥ 0.6	Gemini Avg. Sim	Gemini ≥ 0.6
1	0.332	0/20	0.382	2/20
2	0.306	1/20	0.378	2/20
3	0.415	4/20	0.396	3/20
4	0.367	2/20	0.330	0/20
5	0.390	2/20	0.366	2/20
6	**0.418**	**4/20**	0.422	3/20
7	0.354	2/20	0.393	2/20
8	0.334	0/20	0.430	3/20
9	0.403	2/20	0.380	4/20
10	0.345	2/20	0.378	1/20
11	0.389	3/20	**0.480**	**5/20**
12	0.337	1/20	0.461	4/20

Notably, runs with access to CIDOC CRM relationship data (runs 3, 6, 9, and 12) did not consistently outperform others, and runs provided with the full set of classes and properties tended to underperform. This may suggest that narrowing the input scope to a manually filtered subset, without CIDOC CRM relationship data, helps steer LLM generation toward more meaningful or semantically focused questions.

4.2. Domain Expert Evaluation

For step two, the domain expert evaluation, question sets generated by runs 6 (GPT) and 11 (Gemini), the highest-scoring runs in Table 2, were used. Two domain experts independently expressed their judgment on these question sets, rating each individual question via an online questionnaire[1] on relevance, clarity, and answerability using a 1-5 Likert scale. They were also asked whether any questions in the generated sets could be added to the ground truth set, to select which question set they preferred overall, and to provide free-text comments. The experts were not told which question set was generated by which LLM. The average scores and quantitative analysis of these scores across both experts are summarised in Table 3. The full set of results can be found here [21], and the code used to analyse the results can be found here [23].

While the GPT-generated question set scored slightly higher in average relevance and clarity, the Gemini-generated question set outperformed it in answerability, based on mean ratings across both experts. Inter-rater reliability varied significantly across dimensions and LLMs. For relevance, both models showed extremely low agreement between raters, with intraclass correlation coefficients (ICCs) close to zero or even negative (GPT ICC3k = −0.015, Gemini ICC3k = 0.144), indicating substantial subjectivity in how relevance was interpreted. This is further reflected in the high standard deviations

[1] Full questionnaire: https://forms.gle/SPcFTR8ruxp1DfUD7. This demo version no longer collects responses.

Table 3
Average (Avg) and standard deviation (Std Dev.) of the domain expert (DEx) ratings for the two question sets (Gemini and GPT) across three evaluation dimensions: relevance, clarity, and answerability; the inter-rater reliability for each dimension, measured using intraclass correlation coefficients (ICC 3,k) [27], and the confidence interval (CI) for each ICC.

Metric	Model	DEx1 Avg., Std Dev.	DEx2 Avg., Std Dev.	Overall Avg.	ICC(3,k)	95% CI
Relevance	Gemini	2.55, 1.669	3.2, 1.542	2.875	0.144	[-1.16, 0.66]
Relevance	GPT	3.05, 2.038	3.4, 1.273	3.35	-0.015	[-1.56, 0.60]
Clarity	Gemini	2.95, 1.701	3.75, 1.372	3.1	0.539	[-0.16, 0.82]
Clarity	GPT	3.25, 1.943	3.95, 1.099	3.225	0.839	[0.59, 0.94]
Answerability	Gemini	2.35, 2.159	3.85, 1.268	3.6	0.671	[0.17, 0.87]
Answerability	GPT	1.95, 2.114	3.15, 0.988	2.55	0.839	[0.59, 0.94]

(up to 2.038 for GPT), suggesting that domain experts varied widely in their opinion on how well the questions align with the purpose of a CQ in the context of Tús Maith (relevance). In contrast, GPT achieved strong inter-rater agreement for both clarity (ICC3k = 0.839) and answerability (ICC3k = 0.839), accompanied by comparatively lower standard deviations, indicating more consistent scoring and tighter consensus among experts. Gemini exhibited moderate reliability (clarity ICC3k = 0.539; answerability ICC3k = 0.671) but showed greater variability in scores, especially for answerability, where standard deviations reached 2.159 for domain expert 1. These combined metrics, ICCs capturing rating agreement between the domain experts, the average rating given by them to the generated question sets across the three criteria, and standard deviations indicating score dispersion, highlight the nuanced trade-offs between average performance and consistency of expert assessments, which is critical in qualitative evaluations with subjective criteria.

Expert ratings reflected common patterns: questions rated poorly typically contained ambiguous wording or technical jargon that obscured their intent and reduced accessibility for non-technical users. In contrast, highly rated questions were clear, focused on meaningful historical information, and aligned well with the VRTI KG, making them both understandable and answerable. However, these findings also highlight the need to analyse questions comprehensively across all three dimensions rather than focusing on any single criterion. For example, some questions, such as *"What is the relationship between a group and a person?"*, were rated highly for relevance and clarity by both domain experts but scored lower on answerability due to it referring to information not explicitly contained within the VRTI KG. This underscores that a balance across all dimensions is essential for a question to function well as a template-style CQ. Table 4 presents paired examples of questions across all three evaluation dimensions (Relevance, Clarity, Answerability), showing examples of both the lowest-rated (≤2/5) and highest-rated (≥4/5) questions where both domain experts agreed in their assessments.

Expert comments offer important qualitative insights. While both models generated syntactically valid questions, they were often overly complex or used domain-specific jargon unsuited to non-technical users, for example, *"What type of attribute assignment was carried out?"*. One expert observed that Gemini's questions, although still technical, were easier to interpret and more closely matched the structure and content of the VRTI KG. This likely contributed to both experts ultimately preferring the Gemini-generated set. A notable outcome of the evaluation is that each expert identified at least one Gemini question that could be added to the ground truth CQ set, and one expert did so for GPT. An example of a question a domain expert that found could be added is *"What title does a person have?"*. This indicates that LLM-generated questions, despite the current limitations, can reveal potential gaps in manually constructed question templates.

4.3. Comparative Insights

The comparison between SentenceBERT similarity and expert judgment reveals both alignment and divergence. GPT, despite slightly lower SentenceBERT similarity than Gemini, was rated more relevant

Table 4
Comparison of question quality between Gemini and GPT models across the three evaluation criteria (Relevance, Clarity, Answerability), showing representative examples rated highly and poorly by both domain experts.

Metric	Rating	Gemini Example	GPT Example
Relevance	Poorly Rated	*What type of human-made thing is this?*	*Who carried out the attribute assignment?*
	Highly Rated	*When was a person born?*	*What role did a person have?*
Clarity	Poorly Rated	*What is the preferred identifier of a person?*	*Who is identified by this identifier?*
	Highly Rated	*What is the hectare size of a place?*	*What is the century associated with an event?*
Answerability	Poorly Rated	*What is depicted by a physical thing?*	*What elements compose this object?*
	Highly Rated	*What place is a birth associated with?*	*What is the gender of a person?*

and clear but less answerable. This suggests that it may generate more ambitious or abstract questions. Conversely, Gemini's higher similarity and answerability scores indicate that it may produce questions closer to the current graph content and better suited to near-term user exploration needs. An important confounding factor may be KG class coverage. 7 of 10 (70%) expert ground truth questions were person entity related, whereas only 9 of 20 (45%) Gemini questions and 10 of 20 (50%) GPT questions in their respective best runs referred to people. Future work could investigate whether providing the LLM with class coverage ratios that reflect the most frequent entity types in the KG leads to improved question generation results.

4.4. Summary

Overall, the results highlight that while LLMs can generate semantically rich and partially relevant CQs, there is a trade-off between semantic similarity, readability, and groundedness in the graph. Narrowing input scope and including few-shot examples improved performance, and SentenceBERT evaluation aligned in part with expert assessments. However, inter-rater reliability metrics revealed variation in how experts interpreted key criteria like relevance, underscoring the need for clearer evaluation rubrics for domain experts to follow. While some subjectivity is inevitable, even among domain experts working in the same area provided with a clearer rubric, relevance still remains a valuable criterion as it is one that situates the question within the specific context of template-style CQs, in ways that clarity and answerability alone cannot capture. Further tuning is also needed to ensure the LLM output is suitable for non-expert end users and targeted toward the graph content as it exists currently.

5. Limitations and Future Work

This experiment presents an initial investigation into the use of LLMs for generating template-style CQs to support exploration of the VRTI KG. However, several limitations must be acknowledged.

Firstly, only two LLMs, GPT-4o and Gemini 2.0 Flash, were tested, both using their default settings. As these are commercial models, the prompts were executed via standard API access. Ethical and licensing considerations for such systems are relevant but were not a concern in this case due to the public nature of the VRTI KG. While these are among the most capable publicly available models, future work could explore open-source alternatives such as Mistral [28] or LLaMA 4 [29], which may offer greater customisability or domain-specific fine-tuning. Expanding the range of models evaluated would provide a clearer picture of how model architecture and access conditions influence output quality.

Secondly, the evaluation involved only two domain experts. Although their feedback was detailed and constructive, a larger and more diverse pool of reviewers would allow for a more reliable assessment of the questions' practical relevance, clarity, and answerability. Scaling up this aspect of the study would strengthen the validity of the findings and help identify consistent patterns in expert preference. Additionally, while the clarity criterion, like the others, implicitly includes multiple subcomponents (e.g., grammar, ambiguity, phrasing), no formal rubric was used to disaggregate or define these dimensions. The evaluation instead reflects common humanities practice, where expert interpretation and contextual judgment are often more appropriate than rigid scoring frameworks [30]. Nonetheless, future work may benefit from developing standardised evaluation rubrics to support clearer comparisons across datasets, models, and evaluators.

Thirdly, further analysis needs to be carried out on whether or not the generated questions can be translated into SPARQL without augmentation. Fourthly, while this study focuses on the early stages of the Tús Maith pipeline, namely, the automatic generation and expert review of question templates, it does not yet evaluate how these questions, when instantiated and validated by domain experts, perform when integrated into the VRTI Explorer interface. A natural next step is to assess the effectiveness of these questions in situ, especially in supporting non-technical users' early-stage exploration of the graph. Measuring user engagement, comprehension, and exploratory behaviour in response to these questions could offer valuable insights into their practical impact.

Finally, while the VRTI KG offers a rich and historically grounded test case, generalisability remains an open question. Future work could explore the application of Tús Maith to other types of KGs. Generic KGs such as DBpedia [31] would offer breadth, but a more meaningful test of the framework's relevance would involve similarly complex and semantically rich graphs in the humanities domain. This would better reflect the kinds of exploratory challenges Tús Maith is designed to address.

6. Related Work

A number of recent studies have investigated the use of LLMs in order to generate CQs from KGs, though often with different aims and assumptions than those guiding this work.

RevOnt [32] explores the reverse engineering of CQs from KGs such as Wikidata, drawing on a set of human-authored CQs from existing corpora as ground truth for evaluating output quality. Their definition of a CQ, "a typical query that an expert might want to submit to a knowledge base of its target domain, for a certain task", reflects a focus on ontology modelling and evaluation. In contrast, this experiment defines a CQ as a natural language question that helps convey the scope and potential of a KG to users unfamiliar with its structure, with the aim of supporting exploration rather than formal assessment. Moreover, RevOnt assumes the availability of triple verbalisations, a condition that does not hold for the VRTI KG, which contains over 2.7 million non-verbalised triples and makes the RevOnt pipeline non-transferable to this context.

Other work [33] focuses on generating CQs from KGs constructed from PDF documents. After clustering similar entities and generating summaries, CQs are produced from these summaries and evaluated using zero- and few-shot prompting with cosine similarity scores. However, this introduces a degree of circularity by using the summary to both generate and evaluate the questions. Their work also took an LLM-as-a-judge evaluation approach without the involvement of domain experts, while in contrast, this experiment employs human expert evaluation to assess the quality, relevance, and answerability of generated CQs.

A related line of work by Alharbi et al. spans two papers: the first retrofits CQs to existing ontologies by prompting LLMs to generate multiple questions per triple [34], and the second extends this with an analysis of model settings and prompt design strategies across LLMs [35]. Both papers are motivated by the needs of ontology engineers, aiming to support ontology testing, reuse, and design evaluation rather than exploratory search. While methodologically thorough, both studies are situated within ontology engineering workflows and do not integrate expert-led curation or address the needs of non-technical users. However, their exploration of varied prompt formulations supports the hypothesis also under investigation in this paper, that LLMs may behave differently when cast into specific roles.

While surveys such as [9] comprehensively cover tools for KG exploration, they overlook the initial entry barriers non-technical users face, barriers our work aims to address. Recent benchmarks like Bench4KE [36] focus on automated CQ generation to evaluate ontology quality. In contrast, our approach (1) generates exploratory scaffold CQs tailored for non-technical users, and (2) embeds expert curation to ensure usability. This second point is especially crucial in complex, humanities-based graphs like VRTI, where ontology complexity compounds the Initial Exploration Problem.

Lastly, recent work on CQ benchmarking [37] provides criteria for evaluating CQ quality across syntactic, semantic, and domain-specific dimensions, including SPARQL transformability. While useful as background for designing evaluation strategies, these frameworks are not yet widely adopted and assume the existence of a ground truth CQ set, which, up until this experiment, did not exist for the VRTI KG.

7. Conclusions

This paper has presented Tús Maith as a potential solution to the Initial Exploration Problem, a challenge formally defined here as the set of barriers non-technical users face when first attempting to explore complex KGs. It has also explored the application of LLMs to the generation of CQ templates for a complex historical KG, with the aim of integrating them into the Tús Maith pipeline. KGs such as the VRTI KG, which contains over 2.7 million RDF triples and has been constructed using a variety of often complex ontologies, are typically too large and intricate for domain experts to reasonably recall all relevant classes, properties, and relationships needed to formulate a comprehensive set of template-style CQs manually. In this context, LLMs serve as a cognitive aid, helping experts identify plausible CQ templates that they feel reflect the structure and content of the graph. By combining prompt engineering with structured experimentation across two state-of-the-art LLMs, we have assessed to what extent LLMs can generate CQs to support Tús Maith. Our findings show that, with minimal input, LLMs can generate a promising number of relevant, clear, and answerable questions, particularly when provided with a reduced number of classes and properties and few-shot examples. However, differences in model behaviour, input scope, and prompt design influence output quality. Domain expert evaluation highlighted the potential of LLM-assisted question generation as a curatorial aid, but also underscored the need for expert oversight in validating and refining outputs. Future work will include extending the CQ evaluation using a clearer criteria rubric across a broader pool of experts, performing an end-to-end evaluation of the Tús Maith pipeline in practice, and investigating its generalisability to other similarly complex KGs. Ultimately, this work supports a vision of a semi-automated curation pipeline that blends LLM-generated scaffolding with domain expertise, enabling non-technical end users access to exploration starting points to complex KGs in the humanities domain.

Acknowledgments

The research conducted in this publication was funded by the Irish Research Council under project ID GOIPG/2021/1165 and supported by the Research Ireland ADAPT Centre for Digital Content Technology (grant number 13/RC/2106_P2).

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

[2] D. Lewis, J. Keeney, D. O'Sullivan, S. Guo, Towards a managed extensible control plane for knowledge-based networking, in: R. State, S. van der Meer, D. O'Sullivan, T. Pfeifer (Eds.), Large Scale Management of Distributed Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 98–111.

[3] P. Bhardwaj, J. D. Kelleher, L. Costabello, D. O'Sullivan, Adversarial attacks on knowledge graph embeddings via instance attribution methods, CoRR abs/2111.03120 (2021). URL: https://arxiv.org/abs/2111.03120. arXiv:2111.03120.

[4] RDF, Resource description framework (rdf), http://www.w3.org/RDF/, 2014. [Online; Accessed June 2025].

[5] SPARQL, Sparql query language for rdf, https://www.w3.org/TR/sparql11-query/, 2013. [Online; Accessed June 2025].

[6] H. Li, G. Appleby, C. D. Brumar, R. Chang, A. Suh, Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities, arXiv preprint arXiv:2304.01311 (2023).

[7] B. Kantz, K. Innerebner, P. Waldert, S. Lengauer, E. Lex, T. Schreck, Onset: Ontology and semantic exploration toolkit, 2025. URL: https://arxiv.org/abs/2504.08373. arXiv:2504.08373.

[8] M. Doerr, The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata, AI Mag. 24 (2003) 75–92.

[9] M. Lissandrini, T. B. Pedersen, K. Hose, D. Mottin, Knowledge graph exploration: where are we and where are we going?, SIGWEB Newsl. 2020 (2020). URL: https://doi.org/10.1145/3409481.3409485. doi:10.1145/3409481.3409485.

[10] Virtual Record Treasury of Ireland, Virtual Record Treasury of Ireland, Online resource, 2025. URL: https://virtualtreasury.ie/, online; Accessed June 2025.

[11] Virtual Record Treasury of Ireland, Knowledge graph, Online resource, 2025. URL: https://virtualtreasury.ie/knowledge-graph, online; Accessed June 2025.

[12] R. Battle, D. Kolas, Geosparql: enabling a geospatial semantic web, Semantic Web Journal 3 (2011) 355–370.

[13] L. McKenna, L. Kilgallon, A. Randles, B. Yaman, C. Debruyne, F. Orlandi, G. Munnelly, P. Crooks, D. O'Sullivan, Virtual Record Treasury of Ireland (VRTI) Ontology (v1.2), Web page, 2025. URL: http://ont.virtualtreasury.ie/ontology/index-en.html, revision v1.2, Accessed June 2025.

[14] M. Al-Tawil, V. Dimitrova, D. Thakker, B. Abu-Salih, Emerging exploration strategies of knowledge graphs, IEEE Access 11 (2023) 94713–94731. doi:10.1109/ACCESS.2023.3308514.

[15] E. Kuric, J. Fernández, O. Drozd, Knowledge graph exploration: A usability evaluation of query builders for laypeople, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019, Lecture Notes in Computer Science, Springer, Cham, 2019. doi:10.1007/978-3-030-33220-4_24.

[16] X. Wang, X. Wang, Z. Li, D. Han, Kgnav: A knowledge graph navigational visual query system, Proceedings of the VLDB Endowment 18 (2023). doi:https://doi.org/10.14778/3611540.3611592.

[17] G. K. Q. Monfardini, J. S. Salamon, M. P. Barcellos, Use of competency questions in ontology engineering: A survey, in: J. P. A. Almeida, J. Borbinha, G. Guizzardi, S. Link, J. Zdravkovic (Eds.), Conceptual Modeling, Springer Nature Switzerland, Cham, 2023, pp. 45–64.

[18] OpenAI, Gpt-4o, 2024. URL: https://platform.openai.com/docs/models/gpt-4o, online; Accessed June 2025.

[19] GoogleCloud, Gemini 2.0 flash, 2024. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash, online; Accessed June 2025.

[20] A. Randles, L. McKenna, L. Kilgallon, B. Yaman, P. Crooks, D. O'Sullivan, The knowledge graph explorer for the virtual record treasury of ireland, in: Proceedings of the 9th International Work-

shop on Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs (VOILA 2024), Baltimore, USA, 2024. Co-located with the 23rd International Semantic Web Conference (ISWC 2024), November 11–15, 2024.

[21] C. McNamara, Tús maith - automatic generation of template competency questions, 2025. URL: osf.io/2xyn4.

[22] Y. Rebboud, L. Tailhardat, P. Lisena, R. Troncy, Can llms generate competency questions?, in: The Semantic Web: ESWC 2024 Satellite Events: Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2025, p. 71–80. URL: https://doi.org/10.1007/978-3-031-78952-6_7. doi:10.1007/978-3-031-78952-6_7.

[23] C. McNamara, Tus maith, https://github.com/mcnamacl/Tus-Maith, 2025. Repository containing the code used to generate the template-style CQs, analyse them via SentenceBERT, and analyse the results of the domain expert evaluation.

[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[25] N. Reimers, I. Gurevych, all-mpnet-base-v2, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2021. Online; Accessed June 2025.

[26] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 140 (1932) 1–55.

[27] T. Koo, M. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, Journal of Chiropractic Medicine 15 (2016). doi:10.1016/j.jcm.2016.02.012.

[28] MistralAI, Mistral 7b, https://mistral.ai/news/announcing-mistral-7b, 2023. [Online; Accessed July 2025].

[29] MetaAI, Llama 4, https://ai.meta.com/llama/, 2025. [Online; Accessed July 2025].

[30] J. Edmond, Digital Technology and the Practices of Humanities Research, Open Book Publishers, 2020. URL: https://www.openbookpublishers.com/books/10.11647/obp.0192. doi:10.11647/OBP.0192.

[31] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The Semantic Web (ISWC 2007), volume 4825 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 722–735. URL: https://link.springer.com/chapter/10.1007/978-3-540-76298-0_52. doi:10.1007/978-3-540-76298-0_52.

[32] 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. URL: https://www.sciencedirect.com/science/article/pii/S1570826824000088. doi:https://doi.org/10.1016/j.websem.2024.100822.

[33] D. Di Nuzzo, E. Vakaj, H. Saadany, E. Grishti, N. Mihindukulasooriya, Automated generation of competency questions using large language models and knowledge graphs, in: SEMANTiCS Conference, 2024.

[34] R. Alharbi, V. Tamma, F. Grasso, T. Payne, An experiment in retrofitting competency questions for existing ontologies, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC ’24, ACM, 2024, p. 1650–1658. URL: http://dx.doi.org/10.1145/3605098.3636053. doi:10.1145/3605098.3636053.

[35] R. Alharbi, V. Tamma, F. Grasso, T. R. Payne, The role of generative ai in competency question retrofitting, in: A. Meroño Peñuela, O. Corcho, P. Groth, E. Simperl, V. Tamma, A. G. Nuzzolese, M. Poveda-Villalón, M. Sabou, V. Presutti, I. Celino, A. Revenko, J. Raad, B. Sartini, P. Lisena (Eds.), The Semantic Web: ESWC 2024 Satellite Events, Springer Nature Switzerland, Cham, 2025, pp. 3–13.

[36] A. S. Lippolis, M. D. Ragagni, P. Ciancarini, A. G. Nuzzolese, V. Presutti, Bench4ke: Benchmarking automated competency question generation, 2025. URL: https://arxiv.org/abs/2505.24554. arXiv:2505.24554.

[37] R. Alharbi, J. de Berardinis, F. Grasso, T. Payne, V. Tamma, Characteristics and desiderata for competency question benchmarks, in: The Semantic Web-ISWC, 2024.