Investigating Open Source LLMs to Retrofit Competency Questions in Ontology Engineering

Reham Alharbi[1,2], Valentina Tamma[1], Floriana Grasso[1], Terry R. Payne[1]

[1]University of Liverpool, Liverpool, UK
{R.Alharbi, V.Tamma, F.Grasso, T.R.Payne}@liverpool.ac.uk
[2]Taibah University, Madinah, KSA
rfalharbi@taibahu.edu.sa

Abstract

Competency Questions (CQs) are essential in ontology engineering; they express an ontology's functional requirements as natural language questions, offer crucial insights into an ontology's scope and are pivotal for various tasks, e.g. ontology reuse, testing, requirement specification, and pattern definition. Despite their importance, the practice of publishing CQs alongside ontological artefacts is not commonly adopted. We propose an approach based on Generative AI, specifically Large Language Models (LLMs) for retrofitting CQs from existing ontologies and we investigate how open LLMs (i.e. *Llama-2-70b*, *Mistral 7B* and *Flan-T5-xl*) perform in generating CQs for existing ontologies. We compare these results with our previous efforts using closed-source LLMs and analyse the effect of parameters (e.g. creativity) on the performance. We report high recall and stability of the generated CQs across the systems and we discuss how widespread access to the training data may affect the performance.

Introduction

One of the most time consuming activities in ontology engineering is the elicitation and representation of the requirements scoping the knowledge modelled by an ontology. A common way to represent functional requirements is through the definition of *Competency Questions*, or CQs (Grüninger and Fox 1995) i.e. natural language questions that are limned from a subject domain and from the corresponding tasks that the ontology should support. These CQs are used in different stages of the ontology development lifecycle, including scoping, verification and validation of the knowledge modelled in the ontology (Noy and McGuinness 2001; Poveda-Villalón et al. 2022; Presutti et al. 2009; Sequeda et al. 2019; Suárez-Figueroa, Gómez-Pérez, and Fernández-López 2015). Whilst the formulation of CQs is seen as good practice, some ontology practitioners eschew this step (Alharbi, Tamma, and Grasso 2021; Keet, Mahlaza, and Antia 2019), possibly due to a lack of clear methodologies, guidelines and tools that support the formulation of clear and answerable CQs, and the lack of agreement on what a "good" competency question is. Even when CQs are formulated, they are seldom published as part of the documentation of the ontology, therefore making it difficult to fully understand the scope of the ontology and ultimately hindering its reuse.

Recent advances in transformer architectures, particularly in Large Language Models (LLMs) (Mulla and Gharpure 2023), have sparked the interest of the ontology engineering community, and a number of efforts have emerged that exploit LLMs in different activities in ontology engineering: e.g. for the automatic construction of concept hierarchies (Funk et al. 2023), for learning (Mateiu and Groza 2023; Babaei Giglou, D'Souza, and Auer 2023), and engineering ontologies (Saeedizade and Blomqvist 2024; Fathallah et al. 2024).

In an earlier study (Alharbi et al. 2024), we presented an approach that generates CQs for an existing ontology, to address the lack of published competency questions for a large number of ontologies. Specifically, we investigated the ability of closed-source, or *proprietary* LLMs (i.e. the GPT family of models) to formulate CQs based on triples taken from the original ontology and that mirrored the ones formulated by the developers (*CQ retrofitting*). This initial analysis was conducted across various ontologies from the CORAL repository (Fernández-Izquierdo, Poveda-Villalón, and García-Castro 2019), that publishes the CQs for a number of published ontologies, and confirmed that CQs could not only be generated from statements garnered from an ontology, but that they also closely matched the intended design CQs. In this study we extend the analysis to investigate whether similar results can be obtained using open-source LLM models.

A challenge observed when using LLMs is their non-deterministic behaviour (La Malfa et al. 2023): repeated queries may result in subtly different responses due to a level of stochasticity, which can be controlled through the careful setting of the *creativity (CP)* or *temperature* parameter, resulting in a wider or diverse range of results. Given that the default parameter settings are non-zero, this raises the question as to whether more control can be applied by using a lower value. Together with the definition of different (and increasingly specific) prompts, we explore how varying the *creativity (CP)* can affect the veracity and accuracy of generated CQs within an evaluation using existing benchmark datasets (i.e. ontologies with associated CQs).

Although the analysis presented in this paper extends and, to a certain extent, confirms our previous study (Alharbi

Model	Access Paradigm	Architecture	# Params	Training Data
gpt-3.5 turbo-0613[1]	Proprietary	Decoder-only	20B	Undisclosed
gpt-4-06132[2] (Chang 2023)	Proprietary	Decoder-only	200B	A web-scale corpus
Llama-2-70b (Touvron et al. 2023a)	Open-source	Decoder-only	70B	Publicly available data
Mistral 7B (Jiang et al. 2023)	Open-source	Decoder-only	7B	Undisclosed
Flan-t5-xl (Raffel et al. 2020)	Open-source	Encoder-decoder	3B	Colossal Clean Crawled Corpus (C4)

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo
[2] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

Table 1: Summary of LLMs Including Model Architecture, Number of Parameters, and Training Data.

et al. 2024), it has elicited a number of new insights. The creativity parameter appears to have less effect on the performance of both the open and closed source systems than expected. Furthermore, despite exploring a variety of prompts, a significant overlap of 82% and above was observed across the candidate CQs generated by all prompts within each model and across all models combined, which supports the argument that there is a high degree of replicability and reliability of such evaluations, independent of the creativity parameter's value. The issue of *data leakage* was discussed as a means of explaining the recall results, due to the fact that the original CQs were published together with the ontologies used in the benchmark.

The paper is organised as follows: we first provide a brief overview of the background, followed by the methodology used, including a brief outline of the RETROFIT-CQ pipeline. We then present the results obtained by using open LLMs and we compare with previous results that focussed only on closed models (i.e. the GPT family). We assess the performance of the different LLMs in the RETROFIT-CQ pipeline through precision and recall, and through overlap, i.e the measure of the CQs generated by all systems. We finally examine the effect of creativity on the formulation of CQs from existing ontologies, and we discuss some open issues, e.g. the reliability of the obtained results in view of possible data leakage scenarios.

Background

Large Language Models (LLMs) have shown promise for a plethora of tasks, including the automatic generation of natural language questions (Mulla and Gharpure 2023; Liang et al. 2023) from text or knowledge bases. Auto-regressive LLMs such as those in the GPT family (Ouyang et al. 2022) are deep learning models trained on vast data corpora, and are used to predict the next word in a sequence based on the previous context. Through their use, a new text generation paradigm has emerged whereby a *prompt* guides the generation of various outputs (Mulla and Gharpure 2023). These prompts, consisting of strings prepended to the input context, incorporate control elements (such as keywords) to guide the text generation (Liu et al. 2023). Initial research has already investigated the significant impact of different prompt designs on the performance and outputs of LLMs (Shin et al. 2020), effectively laying the groundwork for the field of prompt engineering (Liu et al. 2023).

Despite the impressive capability that LLMs have to produce syntactically correct and complex natural language, ensuring that this output is meaningful and accurate remains a challenge (Allen, Ilievski, and Joshi 2023). A more nuanced view suggests that LLMs, when combined with traditional symbolic approaches, can play a vital role in knowledge engineering workflows, leading to a new era in knowledge representation that merges explicit and parametric knowledge (AlKhamissi et al. 2022; Allen, Ilievski, and Joshi 2023). The effectiveness of these methods must be validated by addressing LLM-related challenges such as expressivity vs decidability (Pan et al. 2023), thoroughly evaluating approaches that incorporate LLM components (Allen, Stork, and Groth 2023), and tackling issues stemming from insufficient information about LLMs, including their reliability and replicability (La Malfa et al. 2023).

One important distinction when discussing LLMs is between closed source LLMs vs open source LLMs. Closed source LLMs generally follow the Language-Models-as-a-Service (LMaaS) (Sun et al. 2022; La Malfa et al. 2023) model, where LLMs are centrally trained and hosted and, typically, provided on a fee paying licensing subscription or pay-per-use basis. The GPT family by Open-AI[1] is an example of this type of models. By open source models we refer to those systems that have publicly accessible code and architecture, and possibly training dataset, therefore supporting the free use, modification, and distribution.[2] Models belonging to this category include Llama-2-70b, Mistral 7B, and Flan-t5-xl (Table 1). In our previous study (Alharbi et al. 2024) we investigated the retrofitting of competency questions using closed source LLMs. In this paper we extend that study to consider open source LLMs.

Research Methodology

The RETROFIT-CQ approach (Alharbi et al. 2024) addresses the problem of generating candidate CQs for an existing ontology. In principle, the statements in an ontology should be derived from its original CQs. Therefore, the aim of the RETROFIT-CQ pipeline is to reverse this process, by reconstructing the CQs of some ontology, given that ontology's existing statements (expressed as triples). The pipeline consists of three main phases: (i) triples are first extracted from the ontology; (ii) LLM prompts are generated by integrating the triples into a template that includes additional contextual cues (i.e. different prompts with an increasing depth of

[1] https://platform.openai.com
[2] https://www.redhat.com/en/topics/ai/open-source-llm#

context, which are discussed later in this Section); (iii) the prompts are presented to an LLM, resulting in a set of possible competency questions; and (iv) the resulting questions are filtered to remove duplicates and irrelevant questions, resulting in the final set of *candidacy CQs*.

In our previous study (Alharbi et al. 2024), an empirical analysis of the RETROFIT-CQ approach was conducted across closed-source LLMs (*gpt-3.5-turbo-0613* and *gpt-4-06132* - see Table 1). The results supported the claim that LLMs, together with explicit knowledge (in the form of ontology triples) and tailored prompts, can effectively generate valid CQs with a high recall. Furthermore, by varying the degree of contextual content in the prompts, the precision of the resulting CQs could be improved. Subsequent analysis has also suggested that by increasing the *creativity parameter*, additional CQs may be generated.[3]

In this study, we build on these insights to explore whether the choice of LLM model will affect the quality of the candidate CQs, by evaluating open-source LLMs. To do this, we utilise the following three open source models: *Llama-2-70b*, *Mistral 7B*, and *Flan-t5-xl* (summarised in Table 1), in addition to exploring the role that different zero-shot prompts have on generating CQs, as well as observing the influence that different creativity parameter settings have on CQ generation (by comparing deterministic and default values). By restricting the prompting strategy to zero-shot prompts, we are consistent with the first step in the question generation approach by Liang et al. (2023) and generate simple questions that query one-hop relations from the subject of the triples. This is to account for the type of CQs in the datasets used, that mostly contain CQs expressed as simple questions (Wiśniewski et al. 2019).

The open-source models were chosen for this study based on their distinct architectures, with the aim of identifying one from each architectural category and selecting models that had been compared against each other in terms of performance. For example, we selected models from *Llama-2-70b* and *Mistral 7B* for comparison because *Mistral 7B* has demonstrated superior performance across all evaluated benchmarks, outperforming the *Llama-2-13B* model as well as exceeding the capabilities of the larger *LLaMa-34B* model in areas such as mathematics and code generation (Jiang et al. 2023). Furthermore, we used *Flan-t5-xl* for its unique architecture (Encoder-decoder) among all of the chosen models.

We also explore the same hypotheses from our previous study (Alharbi et al. 2024) across open-source LLMs, and contrast these with our previous results:

Hypothesis 1: *Prompting an LLM with more contextual information results in the generation of more concise and coherent responses.*

This hypothesis stems from the premise that additional relevant information could enhance the model's understanding and response accuracy.

[3]The use of different creativity parameters has been explored using the GPT family of models reported at the ESWC 2024 special track on LLMs for KE: https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770001.pdf (to appear).

Hypothesis 2: *Employing the default value of the creativity parameter 'temperature' in an LLM tends to produce responses that are more varied and less focused, in contrast to using a deterministic value which is expected to yield factual responses that are more closely aligned with the original text.*

To facilitate this study, we picked four ontologies with associated CQ datasets; three of which (1–3) were taken from the CORAL repository (Fernández-Izquierdo, Poveda-Villalón, and García-Castro 2019) while the fourth (4) was used in Wiśniewski et al. (2019):

1. *Video Game*;
2. *Dem@care*;
3. *VICINITY Core*; and
4. *African Wildlife*.

The ontologies were randomly selected from those that satisfy the following criteria: (i) the ontologies were produced by different developers (CQ style); (ii) they represent various domains (diversity); and (iii) each had a significant number of published CQs (significance).

Within our experimental methodology, we utilise a number of different prompt templates, that are instantiated using triples taken from the different ontologies. These instantiated prompts are then submitted to the LLMs, resulting in viable CQs, that are subsequently filtered. The use of different prompts allows us to investigate *Hypothesis 1*, as they allow us to examine the impact of transitioning from general to granular when generating candidate CQs; and to understand how LLMs can achieve the highest accuracy in the targeted task. Furthermore, we investigate the effect on the accuracy of the generated CQs of injecting more context to the prompt. However, as highlighted previously (Alharbi et al. 2024) LLMs can generate *narrative questions* which are unsuitable as CQs; i.e. questions that can elicit expansive, descriptive responses (Clandinin 2007), often representing subjective views. For example, if we use the following triple `(:Achievement :isAchievementInGame :Game)` with the prompt template **P1** (described below), then we get the CQ: *"Can you recall an achievement in a game that you found extremely satisfying to unlock"*. Unfortunately, this is not a suitable CQ, and thus should be discarded.

The injection of *context* limits the generation of such questions and ensures that the candidate CQs remain focused on defining the ontology's scope and providing context in terms of *how, where, when, why, who* (Sequeda et al. 2019). We define three prompt templates, each providing increasingly richer context:

P1 *General Competency Questions*: this instructs an LLM to generate competency questions for a given statement: *["Based on <statement>, generate a list of competency questions" avoid using narrative questions + statement]*.

P2 *Definitions of Competency Questions:* this prompt explicitly includes the definition of a CQ: *["Based on the <statement>, generate a list of competency question. Definition of competency questions: the questions that outline the scope of an ontology and provide an idea*

	Prompt	LLMs	Deterministic Creativity (CP=0.0)						Default Creativity (CP=0.7)					
				Number of CQs		Performance				Number of CQs		Performance		
			No. Q.	Candi-date	Valid-ated	Prec.	Rec.	F1	No. Q.	Candi-date	Valid-ated	Prec.	Rec.	F1
Video Game	P1	Llama-2-70b	503	503	90	0.179	0.756	0.289	530	530	118	0.223	0.782	0.347
		Flan-t5-xl	56	56	35	0.625	0.625	0.625	56	56	33	0.589	0.579	0.584
		Mistral 7B	332	332	116	0.349	0.811	0.488	360	360	120	0.333	0.839	0.477
	P2	Llama-2-70b	478	478	114	0.239	0.809	0.368	494	398	106	0.266	0.809	0.401
		Flan-t5-xl	56	56	37	0.661	0.649	0.655	56	56	37	0.661	0.649	0.655
		Mistral 7B	387	387	178	0.460	0.899	0.609	398	398	151	0.379	0.873	0.529
	P3	Llama-2-70b	385	385	174	0.452	0.897	0.601	370	367	173	0.471	0.887	0.616
		Flan-t5-xl	56	56	38	0.679	0.717	0.697	56	56	38	0.679	0.644	0.661
		Mistral 7B	390	390	204	0.523	0.940	0.672	382	374	165	0.441	0.917	0.596
African Wildlife	P1	Llama-2-70b	241	241	103	0.427	1.000	0.599	246	246	100	0.407	0.990	0.576
		Flan-t5-xl	25	24	15	0.625	0.882	0.732	25	22	11	0.500	0.846	0.629
		Mistral 7B	173	155	55	0.355	0.965	0.519	178	172	65	0.378	0.970	0.544
	P2	Llama-2-70b	238	238	106	0.445	**1.000**	0.616	266	266	128	0.481	0.992	0.648
		Flan-t5-xl	25	24	15	0.625	0.882	0.732	25	25	18	0.720	0.900	0.800
		Mistral 7B	168	158	94	0.595	0.990	0.743	191	191	103	0.539	0.990	0.698
	P3	Llama-2-70b	216	215	138	0.642	**1.000**	0.782	209	209	123	0.589	**1.000**	0.741
		Flan-t5-xl	25	24	21	**0.875**	0.955	0.913	25	25	20	**0.800**	0.952	0.870
		Mistral 7B	196	192	116	0.604	0.992	0.751	140	138	82	0.594	0.988	0.742
Dem@care	P1	Llama-2-70b	745	745	163	0.219	0.891	0.351	1387	1199	219	0.183	0.920	0.305
		Flan-t5-xl	141	141	69	0.489	0.697	0.575	145	144	62	0.431	0.705	0.535
		Mistral 7B	684	684	242	0.354	0.924	0.512	953	953	339	0.356	0.963	0.520
	P2	Llama-2-70b	1211	956	206	0.216	0.928	0.350	1222	1101	218	0.198	0.928	0.326
		Flan-t5-xl	144	142	73	0.514	0.785	0.621	144	140	62	0.443	0.756	0.559
		Mistral 7B	908	908	334	0.368	0.971	0.534	883	883	330	0.374	0.968	0.539
	P3	Llama-2-70b	1250	610	254	0.416	0.948	0.579	839	839	185	0.221	0.930	0.357
		Flan-t5-xl	145	141	79	0.560	0.859	0.678	142	141	71	0.504	0.780	0.612
		Mistral 7B	898	898	360	0.401	0.976	0.568	1001	1001	404	0.404	0.978	0.571
Vicinity Core	P1	Llama-2-70b	2244	1873	302	0.161	0.971	0.277	1763	1515	244	0.161	0.961	0.276
		Flan-t5-xl	215	214	77	0.360	0.865	0.508	209	209	60	0.287	0.741	0.414
		Mistral 7B	1437	1437	493	0.343	0.988	0.509	1476	1472	546	0.371	0.993	0.540
	P2	Llama-2-70b	2198	1727	345	0.200	0.983	0.332	2196	1967	328	0.167	0.979	0.285
		Flan-t5-xl	214	214	67	0.313	0.870	0.461	218	218	79	0.362	0.888	0.515
		Mistral 7B	867	866	310	0.358	0.990	0.526	1493	1493	569	0.381	0.997	0.551
	P3	Llama-2-70b	1570	1407	269	0.191	0.982	0.320	2264	2045	357	0.175	0.986	0.297
		Flan-t5-xl	213	213	75	0.352	0.893	0.505	215	214	88	0.411	0.907	0.566
		Mistral 7B	1237	1237	446	0.361	0.996	0.529	1310	866	421	0.486	0.995	0.653

Table 2: Summary of results for different prompts using different Open LLM systems, using the following criteria: number of generated questions (No. Q.), filtered questions in the final output (No. Candidate CQs), number of validated candidate CQs against existing CQs (No. of Validated CQs) and Performance Metrics including Precision, Recall & F1 score. The highest precision and recall results are in bold, while the highest precision and recall per ontology are underlined.

about the knowledge that needs to be entailed in the ontology." avoid using narrative questions + statement].

P3 *Use of a Role with Definitions of Competency Questions:* this contextualises the prompt by specifying the role of "Ontology Engineer", implying a more methodological approach to question formulation that focuses on the structural aspects of the ontology development, with the aim of explicitly generating CQs by including the definition of CQs: *["As an ontology engineer, generate a list of competency questions based on the <statement>. Definition of competency questions: the questions that outline the scope of ontology and provide an idea about the knowledge that needs to be entailed in the ontology" avoid using narrative questions + statement].*

One of the challenges with using LLMs is in their inherent non-deterministic behaviour (La Malfa et al. 2023). Queries can result in subtly different responses, that may vary over time or due to queries that could be posed by other users.[4] Parameters such as the *creativity (CP)* or *temperature* parameter may introduce an additional stochastic element to the generative process, resulting in a wider or diverse range of results. This issue underlies *Hypothesis 2*, which asserts

[4]The fact that prompts may result in a non-deterministic response poses a challenge in any evaluation, as the results will always be indicative, but not necessarily conclusive. However, their value is in the fact that they can provide a good insight into their utility for different domain problems, such as that addressed here within this study.

	Candidate CQs
Llama-2-70b	What are the different types of voice data that can be collected and analyzed?
	How does one differentiate between voice data and other types of audio data?
	What are the main applications of voice data in various industries?
	Can voice data be used to infer emotions or personal characteristics of the speaker?
	What are the ethical considerations when collecting and processing voice data for various purposes?
	How does voice data intersect with other forms of data, such as text or image data?
	Can machine learning algorithms be applied to voice data to improve speech recognition or natural language processing tasks?
	Are there any privacy concerns or regulations that need to be considered when dealing with voice data?
	How can voice data be effectively stored and retrieved from databases or data repositories?
	Are there any standardization efforts or best practices for voice data collection and analysis?
Mistral 7B	What is the definition of VoiceData?
	What are the attributes of VoiceData?
	What are the subclasses of VoiceData?
	What are the relationships between VoiceData and other classes in the ontology?
	What are the properties of VoiceData that distinguish it from other classes in the ontology?
	How does VoiceData relate to the concept of MeasuredData?
	What are the types of VoiceData that can be measured?
	What are the units of measurement for VoiceData?
Flan-t5-xl	What are some other subclasses of MeasuredData?
	What are some other subclasses of MeasuredData?
	What kind of data is included in the category MeasuredData?
	What is the purpose of the measured data in the task?
Bench-mark	**Which are the vocal directed tasks?**
	What is assessed in the sentence repeating task?
	What is assessed in the articulation control task?
	What data are measured for neuromuscular impairment in speech production mechanism?

Table 3: The different *Candidate CQs* generated for the *Dem@care* Ontology Triple `(:VoiceData rdfs:subClassOf :MeasuredData)` with CP=0.0 in LLama2, Mistral and FlanT5. The baseline CQs from the benchmark appear in bold.

that the (non-zero) default values of the *creativity (CP)* result in a greater level of non-determinism than that observed if this parameter is set to zero. Thus we investigate the diversity of text generated by the LLMs by adjusting the creativity (CP) or *temperature* parameter, with the following settings:

- a *deterministic* value of 0.0, which eliminates stochasticity and focuses on the consistent generation of text; and
- the *default* value, that allows the generation of more diverse and creative responses.[5]

In order to validate the candidate CQs generated by our approach, we compare them to the baseline CQs included in the original datasets by using *Sentence-BERT (SBERT)* (Reimers and Gurevych 2019), a modification of the pre-trained BERT network. An advantage of using *SBERT* is that it facilitates the pairwise semantic comparison of sentences through sentence-level embeddings using the cosine similarity metric. By determining a threshold for the cosine similarity, we can determine if two CQs are *similar*, mitigating the effect of paraphrasing, or the use of different morphological structures (e.g. plurals) on the similarity assessment. Thus, a candidate CQ is considered valid if it is sufficiently similar to at least one CQ in the baseline dataset (i.e. the *Existing CQs*, or CQ^E).

The results are reported using the *precision (*$Prec$*)* and *recall (*Rec*)* metrics are based on determining the number of CQs in the relevant original baseline CQ dataset - CQ^E, and the *candidate* CQ set - CQ^C, which correspond to the filtered CQs generated by the LLMs. The metrics used are given below, where $CQ^V \subseteq CQ^C$ is the set of candidate CQs that are assessed as having a similar meaning to those in the baseline CQ (CQ^E) according to *SBERT*, such that the cosine similarity is ≥ 0.7 (i.e. true positives); and the set of unmatched CQs (CQ^U) is the relative complement of the set CQ^V with respect to the set of baseline CQs, such that $CQ^U = CQ^E \setminus CQ^V$ (i.e. the CQs in the baseline set that are not in the set of true positives). The similarity threshold, 0.7, was experimentally determined as the most discriminant, whilst allowing some variance between the questions.

Precision (*Prec*): This is the ratio of the number of *True Positives* ($|CQ^V|$) and the sum of both *True Positives* and *False Positives*, i.e. all of *Candidate CQs* ($|CQ^C|$).

$$Prec. = \frac{|CQ^V|}{|CQ^C|} \tag{1}$$

Recall (*Rec*): Also known as *sensitivity*, this is the ratio of the number of *True Positives* ($|CQ^V|$) and the sum of *True Positives* ($|CQ^V|$) and *False Negatives* ($|CQ^U|$) corresponding to the unmatched CQs in the baseline dataset.

$$Rec. = \frac{|CQ^V|}{|CQ^V| + |CQ^U|} \tag{2}$$

F1 Score (F1): This is the harmonic mean of *Prec* and *Rec*,

[5]The default setting, as of December 2023, was 0.7 but has since been adjusted to 1.0.

	Candidate CQs
CP= 0.7	What are the different types of achievements that can be represented in the ontology? How does the ontology define luck? What are the properties of luck that need to be captured in the ontology? How does luck relate to other concepts in the ontology, such as skill, effort, and opportunity? What are the different ways in which luck can be measured or quantified in the ontology? How does the ontology distinguish between different types of luck, such as good fortune, serendipity, and chance?
CP= 0.0	What is the definition of Luck? What are the properties of Luck? What are the subclasses of Luck? What are the relationships between Luck and other concepts in the ontology? How does Luck relate to the concept of Achievement? What are the different types of Luck? How does Luck affect the achievement of goals? How does Luck relate to other factors that influence success? What are the implications of Luck for decision-making and planning?

Table 4: A list of *Candidate CQs* generated for the triple `(:Luck rdfs:subClassOf :Achievement)` taken from the *Video Game Ontology*, when using P3 with CP=0.0 and CP=0.7 in Mistral 7B.

provides a balanced metric that accounts for both the precision and recall of the results.

$$F1 = 2 \times \frac{Prec \times Rec}{Prec + Rec} \quad (3)$$

Evaluation

The comparative analysis of various LLMs across different ontologies when using the *deterministic* CP value (i.e. CP=0.0) illustrates that these models can generate valid CQs, as evidenced by their high recall results (Table 2). However, their precision varies significantly. *Flan-t5-xl* consistently outperforms other models, demonstrating high precision and competitive recall across most tasks. For example, in the *Video Game* ontology, the best results are obtained with Prompt **P3**, where *Flan-t5-xl* achieves a precision of 0.679 and a recall of 0.717. Similarly, in the *African Wildlife* and *Dem@care* ontologies, *Flan-t5-xl* excels with the same prompt, showing precision values of 0.875 and 0.560, and recall values of 0.955 and 0.859, respectively. Although *Mistral 7B* achieves the highest recall in the *Vicinity Core* ontology at 0.996, *Flan-t5-xl* still offers a balanced performance. These findings suggest that Prompt **P3** generally yields the best results, indicating its effectiveness in extracting relevant information. Similarly analysing the results for the *default* CP value (CP=0.7), *Flan-t5-xl* maintains its superior performance (Table 2). In the *Video Game* and *African Wildlife* ontologies, *Flan-t5-xl* again achieves the highest precision and recall using Prompt **P3**, with precision values of 0.679 and 0.8, and recall values of 0.644 and 0.952, respectively.

Although *Mistral 7B* demonstrates high recall in the *Dem@care* and *Vicinity Core* ontologies, *Flan-t5-xl* maintains a good balance with Prompt **P3**. However, *Llama-2-70b* exhibits the lowest precision compared to the other models. This observation may be related to the nature, style, and length of the CQs generated by *Llama-2-70b*, even with the deterministic setting in **P3**, as opposed to the other models. For example, consider the CQs generated for the triple `(:VoiceData rdfs:subClassOf :MeasuredData)` in the *Dem@care* ontology as illustrated in Table 3. *LLama-2-70b's* CQs focus on practical applications, such as how voice data can be collected and applied across various industries, while Mistral 7B's questions concentrate on defining and categorising voice data within the ontology. *FlanT5's* questions, although relevant, provide a broader view by situating voice data within the larger category of `:MeasuredData` but lack the detailed focus exhibited by *Mistral 7B*, which aligns closely with the existing CQs, offering a coherent and comprehensive understanding. A similar observation was also made by Nadeau et al. (2024), who introduced new datasets to assess the safety of LLMs in enterprise environments, focusing on their adherence to instructions and the production of factual, unbiased, grounded, and appropriate content. They concluded that *Llama-2-70b* excels in factuality and managing biased content but that it tends to produce hallucinated content more frequently than *Mistral 7B*. On the other hand, *Mistral 7B* shows the lowest tendency to hallucinate but struggles with handling toxic content effectively (i.e. generating offensive content despite being instructed not to do so). However, *Mistral 7B* exhibits strong performance when handling a mixture of tasks and safety concerns within specialised, narrow domains (Nadeau et al. 2024).

Prompt **P3** continues to yield the best results across all ontologies, reinforcing its effectiveness. This experiment supports *Hypothesis 1*, which states that context-rich prompts lead to more precise and coherent responses from LLMs. Both experiments with open and closed-source LLMs have demonstrated that embedding context in terms of term definitions and roles effectively extracts relevant information from the models.

However, the findings indicate that the performance of open-source LLMs is more sensitive to changes in the CP, with higher *temperatures* leading to increased variability and potential inaccuracies. For example, in the *Video Game* ontology, *Mistral 7B* generates diverse questions in both CP settings using **P3** given the triple `(:Luck rdfs:subClassOf :Achievement)` as illustrated in Table 4. Conversely, the CQs generated with the *deterministic* CP value (CP=0.0) exhibit a more focused and concise approach, querying the basic definitions, properties, and subclasses directly related to `:Luck`. These questions are highly structured, aiming for clarity and directly relevant responses that are crucial for ontology development. Furthermore, these CQs also explore how `:Luck` interacts with other factors that influence success in games, such as decision-making and planning, thus providing a solid foundation for how `:Luck` might be integrated or utilised within the ontology's framework.

Ontology	Category	Llama-2-70b		Flan-t5-xl		Mistral 7B		All 3 LLms	
		0.0	0.7	0.0	0.7	0.0	0.7	0.0	0.7
Video Game	# Candidate CQs	1339	1390	168	168	1109	1132	2615	2689
	# Overlapping CQs	857	914	107	108	901	958	1976	2097
	Overlapping CQ %	64%	65.76%	63.69%	64.29%	81.24%	84.63%	**75.56%**	**77.98%**
African Wildlife	# Candidate CQs	694	721	72	72	505	501	1271	1294
	# Overlapping CQs	507	532	58	57	456	455	1042	1079
	Overlapping CQ %	73.05%	73.79%	80.56%	79.17%	90.30%	90.82%	**81.98%**	**83.38%**
Dem@care	# Candidate CQs	2813	3139	424	425	2490	2837	5727	6401
	# Overlapping CQs	2169	2385	290	299	2104	2477	4690	5295
	Overlapping CQ %	77.11%	75.98%	68.40%	70.35%	84.50%	87.31%	**81.89%**	**82.72%**
VICINITY Core	# Candidate CQs	5007	5527	641	641	3540	4275	9188	10443
	# Overlapping CQs	3852	4236	314	328	2905	3566	7291	8403
	Overlapping CQ %	76.93%	76.64%	48.99%	51.17%	82.06%	83.42%	**79.35%**	**80.47%**

Table 5: Summary of the overlap in *Candidate CQs* for each ontology, including the total number of *Candidate CQs* (# Candidate CQs), the total number of overlapping CQs, and their percentage (Overlapping CQ %).

Overall, the questions resulting from using the *default* value (CP=0.7) enable a broader, more creative exploration of how `:Luck` could be conceptualised within the ontology, which could be beneficial during the initial stages of ontology development or revisions. In contrast, the questions from the *deterministic* value (CP=0.0) are more precise and geared towards firming up the ontology's existing structure, making them more suitable for refining the ontology to ensure clear and actionable definitions and relationships, or during the definitions of SPARQL based tests.

Therefore, *Hypothesis 2*, which states that more robust and reliable CQs can be obtained by reducing the *creativity parameter (temperature value)*, is supported by the open-source LLMs. Specifically, the comparison between *temperature* values 0.0 and 0.7 reveals that higher temperatures lead to increased variability and potential hallucinations in the models' outputs. At CP=0.0, models such as *Flan-t5-xl* and *Mistral 7B* exhibit more deterministic behaviour with consistent precision and recall. At CP=0.7, the models, especially *Mistral 7B*, show higher recall but increased hallucinations, affecting precision.

Similar to the approach taken with the closed-source LLMs, the overlap among the resulting CQs was examined to verify if similar CQs were generated despite variations in the prompt, CP value, and LLM used. This analysis was conducted primarily to understand the behaviour of open-source LLMs and to address concerns regarding the replicability of their performance. Our findings can be summarised by the following two observations:

- The recall value exhibits only a slight difference when the CP is varied (Table 2), indicating that this parameter minimally impacts the effectiveness of the resulting CQs;
- A consistent pattern in the CQs produced by different prompts is observed in the overlap (Table 5) of the generated CQs, irrespective of the LLM used.

Table 5 presents, for each ontology, the overlap of *Candidate CQs* generated by all prompts in the three LLMs individually and also collectively across all LLMs. The overlap is computed for all prompts under each creativity parameter setting (the highest overlap obtained for each ontology across both creativity parameters is highlighted in bold). For example, the overlap of *Candidate CQs* for the *Video Game* ontology, across all prompts is 64%, 63.69% and 81.24% respectively for *Llama-2-70b*, *Flan-t5-xl* and *Mistral 7B* when CP=0.0 (conversely, 65.76%, 64.29% and 84.63% for CP=0.7). These results highlight the reliability of open-source LLMs in consistently retrofitting CQs across various settings, laying a solid foundation for further exploration into how these models maintain performance stability under varying conditions.

Discussion

The analysis presented in this paper extends and confirms, at least partially, the findings of our previous study (Alharbi et al. 2024), where we focused on closed LLMs (i.e. the GPT family). One observation based on the results obtained is the fact that the creativity parameter seems to have little effect on the performance of both the open and closed source systems. In particular, there is a marginal difference in the recall obtained for all systems, whose statistical significance might be worth investigating in future studies. A potential explanation for this minimal variance is the reported non-determinism of *gpt-3.5-turbo* and *gpt-4*, even at a temperature setting of zero.[6] This suggests that while the temperature control variable is undoubtedly helpful (as noted by La Malfa et al. (2023)), its impact on the performance of RETROFIT-CQ in our experiments appears to be of little significance. Precision seems to be more affected from the change in creativity; however for both closed and open source systems, there is no clear improvement or degradation of performance, both with respect to precision and to recall and for all prompts. Furthermore, we observe a significant overlap among the candidate CQs generated by all prompts within each model and across all models combined, as illustrated in Table 5. Given that we employed zero-shot prompts—crafted without adding ground truth examples or fine-tuning—and still observed an overlap of 75.56% and above, this suggests replicability and reliability of our experiment, independent of the creativity parameter's value.

[6] https://152334h.github.io/blog/non-determinism-in-gpt-4/

	Prompt	LLMs	Deterministic Creativity (CP=0.0)						Default Creativity (CP=0.7)					
				Number of CQs		Performance				Number of CQs		Performance		
			No. Q.	Candidate	Validated	Prec.	Rec.	F1	No. Q.	Candidate	Validated	Prec.	Rec.	F1
Video Game	P1	gpt-3.5-turbo	555	555	251	0.452	0.980	0.619	543	543	348	0.641	0.991	0.779
		gpt-4	776	591	482	0.816	0.998	0.898	1249	1205	963	0.799	**0.999**	0.888
	P2	gpt-3.5-turbo	570	569	399	0.701	0.988	0.820	567	567	365	0.644	0.979	0.777
		gpt-4	1033	810	639	0.789	0.995	0.880	1084	1061	852	0.803	**0.999**	0.890
	P3	gpt-3.5-turbo	570	565	434	**0.844**	0.995	0.914	570	565	429	0.759	0.995	0.861
		gpt-4	1197	911	759	0.833	**0.999**	0.908	797	765	628	**0.821**	0.998	0.901
African Wildlife	P1	gpt-3.5-turbo	215	213	136	0.638	0.986	0.775	206	205	128	0.624	0.992	0.766
		gpt-4	496	373	156	0.418	0.987	0.588	274	266	128	0.481	0.970	0.643
	P2	gpt-3.5-turbo	260	258	151	0.585	0.987	0.735	260	259	141	0.544	0.986	0.701
		gpt-4	423	357	186	0.521	0.989	0.683	441	437	173	0.396	0.989	0.565
	P3	gpt-3.5-turbo	270	256	185	0.723	0.995	0.837	265	262	198	0.756	0.995	0.859
		gpt-4	255	174	94	0.540	0.979	0.696	517	459	229	0.499	0.991	0.664
Dem@care	P1	gpt-3.5-turbo	1360	1339	474	0.354	0.977	0.520	1329	1319	452	0.343	0.985	0.508
		gpt-4	2039	1660	512	0.308	0.987	0.470	2134	2101	552	0.263	0.987	0.415
	P2	gpt-3.5-turbo	1435	1418	403	0.284	0.976	0.440	1428	1406	423	0.301	0.979	0.460
		gpt-4	2574	2042	633	0.310	0.994	0.473	2681	2628	780	0.297	0.989	0.457
	P3	gpt-3.5-turbo	1461	1386	622	0.449	0.995	0.619	1475	1459	616	0.422	0.995	0.593
		gpt-4	2850	2129	656	0.308	0.997	0.471	2929	2811	863	0.307	0.994	0.469
VICINITY Core	P1	gpt-3.5-turbo	2179	2119	501	0.236	0.990	0.382	2177	2160	573	0.265	0.995	0.419
		gpt-4	4320	3428	1122	0.327	0.996	0.493	4444	4276	1430	0.334	**0.999**	0.501
	P2	gpt-3.5-turbo	2219	2150	547	0.254	0.993	0.405	2202	2199	596	0.271	0.998	0.426
		gpt-4	4549	3505	1333	0.380	**0.999**	0.550	4824	4695	1723	0.367	**0.999**	0.537
	P3	gpt-3.5-turbo	2249	2115	947	0.448	**0.999**	0.618	2265	2230	947	0.425	**0.999**	0.596
		gpt-4	4958	3863	1485	0.384	**0.999**	0.555	4975	4787	1887	0.623	**0.999**	0.767

Table 6: Summary of results for different prompts using different Closed LLM systems (Alharbi et al. 2024), using the following criteria: number of generated questions (No. Q.), filtered questions in the final output (No. Candidate CQs), number of validated candidate CQs against existing CQs (No. of Validated CQs) and Performance Metrics including Precision, Recall & F1 score. The highest precision and recall results are in bold, while the highest precision and recall per ontology are underlined.

One aspect to consider, especially when considering recall, is a possible *data leakage* or *data contamination* (La Malfa et al. 2023) effect due to the fact that the original CQs were published, sometimes together with the ontologies used in the benchmark. This might explain the almost perfect recall observed for Chat GPT 3.5 and 4.0, where we notice that the recall for both closed source LLMs is very high, almost close to 1.000. One hypothesis for this behaviour is the lack of precise information regarding the training data. For the GPT models this is thought to be documents published on the web and data licensed from third-party providers at different cutoff points (September 2021 vs December 2023). All of the ontologies included in the benchmarks were published before 2021, therefore it is likely that the competency questions were included in the training data and are retrieved in the set of the true positives; this is something that should be investigated further. However, this does not seem to be the case for the open source LLMs, where recall varies between 0.579 (for Prompt **P1** using *Flan-t5-xl* on the Video Game ontology) and 1.000. In running this evaluation with the chosen open source LLMs, we opted to use the pre-trained models with no fine-tuning, to have comparable experimental settings to our previous experiments conducted on *GPT3.5* and *GPT4*. Analogously to the GPT systems, there is no clear knowledge about the training data used by the different open LLM systems. For example, the research paper that introduces *Llama-2-70b* (Touvron et al. 2023b) omits details on specific data sources, but states that *Llama-2-70b* was trained with 2 trillion tokens (i.e. the numerical representation of words, word parts, phrases and other semantic fragments that transformer-based neural networks use for language processing) from publicly available sources. *Flan T5* uses the *Colossal Clean Crawled Corpus (C4)*, containing text and code scraped from the internet (Raffel et al. 2020), and no information is available regarding Mistral, according to the launch blog site.[7] It is plausible that all the systems used publicly available web documents, but for these systems the results suggest that the data leakage effect is less evident.

This lack of precise information on data used in pretraining makes it impossible to identify specific documents relevant to each ontology, which could be used in future for fine tuning. However, we can hypothesise that some design CQs for both the *Dem@care* and *VICINITY Core* ontologies, which are multidisciplinary domains, likely contain long-tail knowledge. This complexity is further complicated by the LLMs' apparent difficulty in generating fact-based questions effectively, suggested by the generation of irrelevant

[7] https://mistral.ai/news/announcing-mistral-7b/

	Prompt	Number and percentage of unmatched existing CQs											
		Llama-2-70b				Flan-t5-xl				Mistral 7B			
		CP=0.0		CP=0.7		CP=0.0		CP=0.7		CP=0.0		CP=0.7	
Video Game	P1	(29)	43.94%	(33)	50%	(21)	31.82%	(24)	36.36%	(27)	40.91%	(23)	34.85%
Existing CQs: 66	P2	(27)	40.91%	(25)	37.88%	(20)	30.30%	(20)	30.30%	(20)	30.30%	(22)	33.33%
# of Triples: 57	P3	(20)	30.30%	(22)	33.33%	(15)	22.73%	(21)	31.82%	(13)	19.70%	(15)	22.73%
African Wildlife	P1	(0)	N/A	(1)	7.14%	(2)	14.29%	(2)	14.29%	(2)	14.29%	(2)	14.29%
Existing CQs: 14	P2	(0)	N/A	(1)	7.14%	(2)	14.29%	(1)	7.14%	(1)	7.14%	(1)	7.14%
# of Triples: 26	P3	(0)	N/A	(0)	N/A	(1)	7.14%	(2)	14.29%	(1)	7.14%	(1)	7.14%
Dem@care	P1	(20)	18.69%	(19)	17.76%	(30)	28.04%	(26)	24.30%	(20)	18.69%	(13)	12.15%
Existing CQs: 107	P2	(16)	14.95%	(17)	15.89%	(20)	18.69%	(20)	18.69%	(10)	9.35%	(11)	10.28%
# of Triples: 146	P3	(14)	13.08%	(14)	13.08%	(13)	12.15%	(20)	18.69%	(9)	8.41%	(9)	8.41%
VICINITY Core	P1	(9)	15.79%	(10)	17.54%	(12)	21.05%	(21)	36.84%	(6)	10.53%	(4)	7.02%
Existing CQs: 57	P2	(6)	10.53%	(7)	12.28%	(10)	17.54%	(9)	15.79%	(3)	5.26%	(2)	3.51%
# of Triples: 226	P3	(5)	8.77%	(5)	8.77%	(9)	15.79%	(10)	17.54%	(2)	3.51%	(2)	3.51%

Table 7: Number (percent) of unmatched Existing CQs for each prompt and LLM, comparing deterministic (CP=0.0) and default (CP=0.7) values.

candidate CQs, even when a deterministic value is set for the creativity parameter (and illustrated in Table 2 showing the difference between Candidate CQs and Validated CQs).

Variability is also noticeable across the systems and by ontology; Table 6 reports the corresponding data for the GPT models as published in our ESWC 2024 special track on LLMs for KE paper.[3]

The best performance (in terms of precision and recall) for open source models is achieved over the African Wildlife Ontology, whereas the best performance for closed source models is achieved with the Video Game ontology.

In both open and closed source models, prompt paraphrasing significantly impacts reasoning, with Prompt **P3** resulting in the best performance overall, indicating that providing context (in the form of Triples), method (through the role of an Ontology Engineer), and term definitions (as in CQs definitions) within prompts leads to the highest precision on the target task.

One further issue pertains the generation of CQs that are not matched in the benchmark (Tables 7 and 8). Closed source models tend to generate a smaller number of CQs that are not matched by the benchmark. Some of these correspond to competency questions identified in the benchmark, but not reflected in the ontology (Alharbi et al. 2024). The number of unmatched CQs, however becomes significant within open source LLMs for all ontologies, with the exception of the African Wildlife ontology. Although a closer analysis of these unmatched CQs is currently ongoing, a possible hypothesis is that the African Wildlife ontology is explicitly used in tutorials and experiment on the formulation of CQs, and therefore might have affected the way these have been used in the training of the models, suggesting that the GPT models are more sensitive to this type of data.

An interesting issue, which needs to be investigated further, is a close analysis of the questions generated. As it was the case for the closed source models, LLMs tend to generate a large number of candidate CQs, therefore degrading the precision. However, a domain expert validation of these CQs should be performed, to identify those cases where these are CQs that were considered during the ontology construction process and then omitted, or discarded. In our previous work (Alharbi et al. 2024) we conducted a small study on a Solar System Ontology developed by a member of our group and whose competency questions have not been published online, hence eliminating any risk of data leakage due to the training data set including these CQs already. We ran RETROFIT-CQ over the produced ontology and asked the ontology developer to validate the generated questions. The interview conducted with the developer confirmed that the CQs generated by the pipeline accurately captured the initial requirements (with a precision over 75%) and that all the intended original CQs were matched. Interestingly, the developer also identified a number of additional CQs that in hindsight captured some of the intended meaning of the ontology and could have been included in the requirement elicitation phase. Some of these were not included in the existing CQs but captured the knowledge modelled by the ontology. Indeed, ontology developers aim to identify CQs that are representative of the requirements of the ontology, however there is no guarantee that these are exhaustive or comprehensive. Therefore, by retrofitting CQs at a later stage, additional CQs can emerge that can be useful when evaluating the ontology design by: 1) translating CQs into SPARQL queries; and 2) querying the populated ontology to stress test the ontology design process and anticipate unintended uses of the ontology.

A final issue that warrants more attention is the fine-tuning of models. In principle, with open source models in particular, we have the potential to train the models over specific datasets that should improve the performance of competency question generation. However, understanding the characteristics of this dataset is an open problem. One obvious way would be to train the LLM over domain specific text documents, similarly to the approach by Antia and Keet (2023). However, this would also require the identification of suitable documents to include in the training set, and solid governance approaches to keep the training set up to date. Furthermore, task specific benchmarks could be designed as a community effort to improve and stress test the performance of LLMs when used to generate or retrofit CQs.

	Prompt	Unmatched CQs (#) %							
		gpt-3.5-turbo				gpt-4			
		CP=0.0		CP=0.7		CP=0.0		CP=0.7	
Video Game	P1	(5)	7.57%	(3)	4.54%	(1)	1.51%	(1)	1.51%
Existing CQs: 66	P2	(5)	7.57%	(8)	12.12%	(3)	4.54%	(1)	1.51%
# of Triples: 57	P3	(2)	3.03%	(2)	3.03%	(1)	1.51%	(1)	1.51%
African Wildlife	P1	(2)	14.28%	(1)	7.14%	(2)	14.28%	(4)	28.57%
Existing CQs: 14	P2	(2)	14.28%	(2)	14.28%	(2)	14.28%	(2)	14.28%
# of Triples: 26	P3	(1)	7.14%	(1)	7.14%	(2)	14.28%	(2)	14.28%
Dem@care	P1	(11)	10.28%	(7)	6.54%	(7)	6.54%	(7)	6.54%
Existing CQs: 107	P2	(10)	9.34%	(9)	8.41%	(4)	3.73%	(9)	8.41%
# of Triples: 146	P3	(3)	2.80%	(3)	2.80%	(2)	1.86%	(5)	4.67%
VICINITY Core	P1	(5)	8.77%	(3)	5.26%	(4)	7.01%	(2)	3.50%
Existing CQs: 57	P2	(4)	7.01%	(1)	1.75%	(2)	3.50%	(2)	3.50%
# of Triples: 226	P3	(1)	1.75%	(1)	1.75%	(1)	1.75%	(1)	1.75%

Table 8: Number and % of unmatched Existing CQs for each prompt and LLM for different creativity parameter values.

Conclusions

This paper investigates whether Large Language Models can be used to retrofit competency questions to existing ontologies. In this study we focus our attention to open source LLMs (i.e. *Llama-2-70b*, *Mistral 7B* and *Flan-T5-xl*) with the aim of complementing and expanding our previous work that focussed on closed models (Alharbi et al. 2024). The results obtained in this study confirm that LLMs can support the generation of competency questions over existing ontologies and the high recall values in the experiments over benchmark ontologies (with their corresponding existing competency questions) confirm that the generated questions match those that were used during the ontology construction process to elicit requirements. However, the values for recall in open source models are more variable, ranging from 0.579 to 1.000. This might be due to the difference in training data used by the various systems but also to the difference in models between the different architectures. Notably, varying the temperature parameter from its default value to its lower bound seems to reduce the number of *hallucinations* in the form of irrelevant or non factually relevant competency questions; however this has only marginal effect on the values of precision and recall.

Acknowledgments

We are grateful to Jacopo de Berardinis for valuable discussions on the results of this work.

References

Alharbi, R.; Tamma, V.; and Grasso, F. 2021. Characterising the Gap Between Theory and Practice of Ontology Reuse. In *Proceedings of the 11th Conference on Knowledge Capture, K-CAP 2021*, 217–224.

Alharbi, R.; Tamma, V.; Grasso, F.; and Payne, T. 2024. An Experiment in Retrofitting Competency Questions for Existing Ontologies. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, 1650–1658.

AlKhamissi, B.; Li, M.; Celikyilmaz, A.; Diab, M. T.; and Ghazvininejad, M. 2022. A Review on Language Models as Knowledge Bases. arXiv:2204.06031.

Allen, B. P.; Ilievski, F.; and Joshi, S. 2023. Identifying and Consolidating Knowledge Engineering Requirements. arXiv:2306.15124.

Allen, B. P.; Stork, L.; and Groth, P. 2023. Knowledge Engineering Using Large Language Models. *Transactions on Graph Data and Knowledge*, 1(1): 3:1–3:19.

Antia, M.-J.; and Keet, C. M. 2023. Automating the Generation of Competency Questions for Ontologies with AgOCQs. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, 213–227. Springer.

Babaei Giglou, H.; D'Souza, J.; and Auer, S. 2023. LLMs4OL: Large Language Models for Ontology Learning. In *The Semantic Web – ISWC 2023*, 408–427. Springer Nature Switzerland.

Chang, E. Y. 2023. Examining gpt-4: Capabilities, implications and future directions. In *The 10th International Conf. on Computational Science and Computational Intelligence*.

Clandinin, D. 2007. *Handbook of Narrative Inquiry: Mapping a Methodology*. Thousand Oaks, California: SAGE Publications, Inc.

Fathallah, N.; Das, A.; De Giorgis, S.; Poltronieri, A.; Haase, P.; and Kovriguina, L. 2024. NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning. In *Extended Semantic Web Conference, ESWC2024*. Hersonissos, Greece.

Fernández-Izquierdo, A.; Poveda-Villalón, M.; and García-Castro, R. 2019. CORAL: A Corpus of Ontological Requirements Annotated with Lexico-Syntactic Patterns. In *Proc. of the 16th International Conf. on The Semantic Web, ESWC 2019*, 443–458.

Funk, M.; Hosemann, S.; Jung, J. C.; and Lutz, C. 2023. Towards Ontology Construction with Language Models. In *Proceedings of the KBC-LM'23: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC*. CEUR Workshop Proceedings.

Grüninger, M.; and Fox, M. S. 1995. The Role of Competency Questions in Enterprise Engineering. In Rolstadås, A., ed., *Benchmarking — Theory and Practice*, 22–31. Springer.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; et al. 2023. Mistral 7B. arXiv:2310.06825.

Keet, C. M.; Mahlaza, Z.; and Antia, M.-J. 2019. CLaRO: a Controlled Language for Authoring Competency Questions. In Garoufallou, E.; Fallucchi, F.; and William De Luca, E., eds., *Metadata and Semantics Research*, 3–15.

La Malfa, E.; Petrov, A.; Frieder, S.; Weinhuber, C.; Burnell, R.; Nazar, R.; et al. 2023. Language Models as a Service: Overview of a New Paradigm and its Challenges. arXiv:2309.16573.

Liang, Y.; Wang, J.; Zhu, H.; Wang, L.; Qian, W.; and Lan, Y. 2023. Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4329–4343. Association for Computational Linguistics.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9).

Mateiu, P.; and Groza, A. 2023. Ontology engineering with Large Language Models. In *Proceedings of the 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 226–229.

Mulla, N.; and Gharpure, P. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1): 1–32.

Nadeau, D.; Kroutikov, M.; McNeil, K.; and Baribeau, S. 2024. Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations. arXiv:2404.09785.

Noy, N. F.; and McGuinness, D. L. 2001. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford knowledge systems laboratory technical report KSL-01-05.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; et al. 2022. Training language models to follow instructions with human feedback. In *Proc. of the Advances in Neural Information Processing Systems, NeurIPS 2022*, volume 35, 27730–27744.

Pan, J. Z.; Razniewski, S.; Kalo, J.-C.; Singhania, S.; Chen, J.; Dietze, S.; et al. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1): 2:1–2:38.

Poveda-Villalón, M.; Fernández-Izquierdo, A.; Fernández-López, M.; and García-Castro, R. 2022. LOT: An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111: 104755.

Presutti, V.; Daga, E.; Gangemi, A.; and Blomqvist, E. 2009. EXtreme Design with Content Ontology Design Patterns. In *Proc. of the 2009 International Conf. on Ontology Patterns*, volume 516, 83–97. CEUR-WS.org.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Reimers, N.; and Gurevych, I. 2019. 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)*, 3982–3992.

Saeedizade, M. J.; and Blomqvist, E. 2024. Navigating Ontology Development with Large Language Models. In *The Semantic Web*, 143–161. Springer Nature Switzerland.

Sequeda, J. F.; Briggs, W. J.; Miranker, D. P.; and Heideman, W. P. 2019. A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases. In *Proceedings of the 18th International Semantic Web Conference, ISWC 2019*, 526–545. Springer International Publishing.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 4222–4235.

Suárez-Figueroa, M. C.; Gómez-Pérez, A.; and Fernández-López, M. 2015. The NeOn Methodology framework: A scenario-based methodology for ontology development. *Applied ontology*, 10(2): 107–145.

Sun, T.; Shao, Y.; Qian, H.; Huang, X.; and Qiu, X. 2022. Black-Box Tuning for Language-Model-as-a-Service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 20841–20855. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; et al. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wiśniewski, D.; Potoniec, J.; Ławrynowicz, A.; and Keet, C. M. 2019. Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL. *Journal of Web Semantics*, 59: 100534.