Large Language Models as Knowledge Evaluation Agents

George Hannah[1,*], Jacopo de Berardinis[1], Terry R. Payne[1], Valentina Tamma[1], Andrew Mitchell[2], Ellen Piercy[2], Ewan Johnson[2], Andrew Ng[2], Harry Rostron[2] and Boris Konev[1]

[1]Department of Computer Science, University of Liverpool, Brownlow Hill, Liverpool, L69 7ZX, United Kingdom
[2]Materials Innovation Factory, University of Liverpool, 51 Oxford Street, Liverpool, L7 3NY, United Kingdom

Abstract

With the recent rise of large language models, there has been a growing interest in employing LLMs in different knowledge engineering tasks, such as supporting the semi-automatic creation of KG schemata. This is the case also when leveraging semi-structured data by translating XML schemata into ontologies, where there is a need to inject additional knowledge that makes explicit the relationships between different entities. We investigate the viability of using LLMs as knowledge evaluation agents to assess the suitability of the injected knowledge; by using Gemini as an LLM-based proxy for a human evaluator, and we configure it with different parameter settings and prompt structures. The responses (i.e. the LLM-based assessment of the relationships) are compared against the set of assessments or evaluations carried out by domain experts. We find that despite encountering some issues, the use of LLMs as a proxy expert shows promise in their ability to understand complex domains and evaluate relationships with respect to that domain.

Keywords

LLM Evaluation, Prompt Engineering, Ontology Engineering

1. Introduction

In recent years there has been a critical increase in the volume of data produced, consumed and stored by different organisations [1]. A significant part of this data is represented in semi-structured data formats such as Extensible Markup Language (XML) and its corresponding schema, and used to annotate data produced by autonomous processes and sensors. The *Analytical Information Markup Language (AnIML)*[1] is one such example used in the domain of autonomous digital chemistry, to represent data produced or consumed by robotic laboratory systems in the configuration and execution of analytical chemistry experiments. Unilever is one organisation that uses AnIML to mark up the data generated in the formulation of experiments executed in different laboratories across its sites, thereby ensuring consistency and the comparability of the chemical formulation of its brand products across the world.

The use of AnIML guarantees syntactic interoperability amongst the different data sources; however whilst AnIML can ensure that documents are structured consistently, it does not ensure that the *meaning* of the elements in the schema is consistently interpreted by the data engineers using it to annotate their data. This is particularly crucial when it comes to understanding the relationships existing between these elements: for example, although the concepts `SampleSet` and `Sample` share a relationship in AnIML, there are two alternate interpretations of this relationship that could be formed: 1) that the `SampleSet` has a member which is a `Sample`; or 2) a `Sample` could be generated from a greater `SampleSet`. The problem is that the true semantics of this relationship is not specified in the structure of the AnIML schema itself.

ELMKE: Evaluation of Language Models in Knowledge Engineering, 2nd Workshop co-located with ESWC-25, Portorož, Slovenia
*Corresponding author.
g.t.hannah@liverpool.ac.uk (G. Hannah); Jacopo.De-Berardinis@liverpool.ac.uk (J. d. Berardinis); trp@liverpool.ac.uk (T. R. Payne); valli@liverpool.ac.uk (V. Tamma); Andrew.Mitchell@unilever.com (A. Mitchell); ellen.piercy@unilever.com (E. Piercy); ewan.johnson@unilever.com (E. Johnson); andrew.ng@unilever.com (A. Ng); harry.rostron@unilever.com (H. Rostron); konev@liverpool.ac.uk (B. Konev)
0000-0002-3218-4559 (G. Hannah); 0000-0001-6770-1969 (J. d. Berardinis); 0000-0002-0106-8731 (T. R. Payne); 0000-0002-1320-610X (V. Tamma); 0000-0002-6507-0494 (B. Konev)

[1]https://www.animl.org/current-schema

Knowledge graphs [2] provide an explicit and machine readable representation of knowledge in an interlinked and structured manner; they are multi-relational graphs composed of entities (nodes) and relations (edges) that represent facts about a given domain or application, often expressed in RDF (Resource Description Framework).[2] KGs are typically modelled according to an explicit schema, or *ontology*, that defines the nature of these relationships and concepts and any constraints on their use (typically represented using RDF Schema[3] or the Web Ontology Language (OWL) [3]). These graphs can then be queried, granting access to advanced data analysis techniques [4]. Thus, the KG schemata define the high-level global semantics and provide a vocabulary for user queries over already existing data sources, therefore providing an alternative, semantically enriched perspective over AnIML data.

However, the process of creating KGs, especially with domains that contain specialised knowledge such as the autonomous chemistry one, can be complex and time consuming due to the fact that it involves injecting precise, and often domain specific knowledge into the schema, which requires significant levels of human effort and expertise.

The use of *Large Language Models (LLM)* such as ChatGPT[4] for ontology engineering has recently garnered significant interest as the large corpus of training data used by these models allows them to mimic humans and produce responses akin to what would be produced by a *Domain Expert (DE)*. In this study, we have used an LLM (ChatGPT) to identify and refine a simple set of labels generated to describe relationships extracted from the AnIML XML schema, as an alternative step in the semi-automatic ontology construction pipeline proposed in Hannah et. al. [5]. We therefore investigate the efficacy of using a different LLM (in this case, Gemini [6]) as an evaluation or proxy agent (thus reducing the dependency of using domain experts) to assess the quality of the relationship labels generated by an LLM. In particular, this study addresses the following research questions (RQs):

- **RQ 1:** How closely can an LLM's evaluation of a label match a DE's evaluation of the same label?
- **RQ 2:** Does providing additional context to the LLM improve its ability to accurately evaluate labels?
- **RQ 3:** What issues present themselves when using LLMs to evaluate the quality of labels for use in ontology engineering?

The remainder of this paper is structured as follows: Section 2 provides an overview of related work, Section 3 describes our experimental setup and how we evaluate the generated results, Section 4 shows the results gleaned from the experiments as well as a discussion into the potential causes for and implications of these results, and in Section 5, the results and contribution of this work are summarised.

2. Related Work

A number of studies have explored approaches to translate semi-structured data formats into a structured, formal representation.[5] Bohring and Auer [8] proposed one approach to translate XML schema directly into an OWL ontology via a mapping based pipeline, by identifying structural patterns defining elements and their relationships in the XML schema, and mapping them to the corresponding OWL pattern. Each relationship is classed as either a *partOf* or a *subClassOf* relation, and labels for these relationships are formed by attaching either the "*has*" or "*dtp*" prefix to the textual label of the entity acting as the range of the property (this is similar to the approach adopted in our earlier work [5]). The OWLMAP approach [9] attempts to represent the hierarchical tree structure of the XML schema in OWL, although this assumes that the hierarchical relationships between elements in an XML schema are used consistently throughout. In practice, the nature of the relationships between concepts in an XML schema are ambiguous and require a *Domain Expert (DE)* to attempt to estimate the intention of the schema author.

[2] https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
[3] http://www.w3.org/TR/rdf11-schema/
[4] https://openai.com/research/gpt-4
[5] A good survey of different approaches for translating semi-structured data formats into OWL can be found in [7].

The requirement of a DE in the translation of XML schema into an ontology brings this problem closer to a classical ontology engineering problem. The pay-as-you-go methodology [10] is one ontology engineering approach that has significant potential for the translation of XML schema into ontologies. It focusses on an initial set of business-questions which are subsequently used to build an ontology that can answer them. The business use-case used for the ontology is then expanded, resulting in a new set of questions that support the further development of the ontology. Where it is not possible for these new questions to be answered by the ontology, the ontology itself is iteratively expanded to cover them. This approach translates well to the context of our use of AnIML as only a fragment of the AnIML schema is used currently. Therefore, focussing efforts in creating a high-quality representation of that schema fragment and then expanding on it in the future offers a promising way of limiting the complexity of the task.

With the rise in popularity of LLMs since the release of ChatGPT-3 in late 2022 [11], there have been several investigations into the viability of utilising LLMs as part of the ontology engineering process [12, 13, 14, 15, 16, 17, 18]. Due to the large volume of training data corpus used in their creation, LLMs have a large breadth of knowledge across many domains allowing them to reduce the workload of the DE and synthesize knowledge that is external to any provided context. Thus, there is the strong potential for their successful deployment within knowledge engineering tasks such as KG completion [19]. In addition to this, LLMs have demonstrated significant success in understanding and processing the meaning of information represented in natural language. This has lead to them being employed to extract facts from large bodies of text, and represent them as triples in a KG [20].

One clear disadvantage of LLMs in these situations, is that LLMs are prone to hallucinations (i.e. the generation of false information [21]). To ensure that these hallucinations are identified early in the ontology engineering process, the quality of the generated knowledge needs to be evaluated. However, this comes at a cost, due to the fact that as the size of the ontology expands, the time required by DEs to evaluate the LLM-generated knowledge also increases. To mitigate this, the use of the LLM-as-a-judge approaches have emerged [22, 23, 24], whereby an LLM is also used to evaluate the suitability, or correctness of a separate LLM-generated response. Whilst these approaches show promising performance, limitations are also present in these methods with regards to the LLM's understanding of more complex, domain specific knowledge [25, 26], thus it may be necessary to ascertain the veracity of such an approach for different problem domains.

3. Method

To explore the viability of using LLM-based agents as a proxy for DEs, and hence investigate the research questions stated in Section 1, an experimental workflow has been defined that evaluates the correctness or appropriateness of a relationship label generated by another LLM. In this case, we define a relationship as either a parent-child, or element-attribute link extracted from an XML schema and a relationship label is a label generated by an LLM to describe the relationship between two concepts.

3.1. The choice of Large Language Models (LLM) for different roles

Within this investigation, LLMs are used to fulfil two separate roles: generating the relationship labels; and assessing the validity of the generated labels. For the first role, the relationship labels were generated using OpenAI's *ChatGPT-4o* with the temperature parameter set to the default setting of 0.7 (thus offering a balance between creativity and consistency). The prompt sent to this model includes: 1) a specific role for the LLM; 2) the context regarding the source of the concepts and the use case of the relationship label; and 3) any documentation present in the schema for each concept.

Label Generation Prompt (ChatGPT-4o): *"You are an analytical chemist. Given a relationship between two concepts linked in an XML schema and defined with the following documentation, this relationship has a simple label describing it and will be used in an ontology representing the domain of chemical formulation experiments being carried out by robots in a lab. Does this*

label accurately represent this relationship? If it doesn't, refine the label to more accurately represent the relationship. If it does, return the original label"

A separate LLM was used for the evaluation (Google's *Gemini-flash-2.0* [6]) for two specific reasons. The first reason is to attempt to mitigate any potential bias that may arise from having the same LLM generate and then evaluate its own responses. The second, is due to capabilities of Gemini's API, which allows the user the ability to force the LLM to fit its response to a specified structure (in this case, an *enumeration*) thus allowing for a consistent, predictable output from the LLM.

In order to assess RQ 2, we require a framework that allows the augmentation of a *base prompt* with different levels of context. The base prompt is defined in such a way as to include the primary instruction that the LLM should follow, but otherwise it lacks any other significant context:

Base Prompt (Gemini-flash-2.0): *"On a scale between 0% and 100%, is the label 'x' an appropriate label to describe the relationship between domain: 'y' and range: 'z'?"*

Given the hypothesis that providing the LLM with more context surrounding the task will result in an increase in accuracy of the resulting responses, three different extensions to the prompty were investigated as part of the empirical study, and compared to the base prompt performance. Each extension included additional contextual information regarding the task:

Definition: This level of context added is with respect to the definition of a relationship, which is a modified version of the definition of a property as stated in the *Resource Description Framework (RDF) Model and Syntax Specification*,[6]. The definition used to augment the base prompt is given below:

A1: *"...A relationship is a specific characteristic describing the connection between two concepts. Each relationship has a specific meaning defining the types of concepts it can connect..."*

Document and use case: This context provides the LLM with: 1) the domain of the origin of the concepts, an XML schema describing chemical formulation experiments carried out by robots; and 2) details of the use case of the relationship label, as well as the ontology representing that same domain.

A2: *"...These concepts are derived from an XML schema describing the domain of chemical formulation experiments carried out by robots in laboratories. The relationship linking these concepts will be used in an ontology representing the same domain..."*

Document: This context provides any of the documentation surrounding the concepts in the relationship that were present in the AnIML schema.

A3: *"...where they are described by the following documentation, Domain: 'domain documentation', and Range: 'range documentation'..."*

An additional component of the evaluation prompt is the definition of the structure of the response. One LLM-based weakness observed during the development of this prompting framework was in its *stochasticity*; i.e. the lack of consistency noted across different, independent LLM responses given the same prompt. It was often noted that when the LLM generates a score for the same relationship label in two different sessions, the responses differed in value (even though they were often clustered in value). Therefore by grouping scores into ranges for evaluations, we aim to mitigate the fuzziness inherent to LLMs. This was addressed within the evaluation by providing the LLM with an enumerated, structured response format, in the form of a *Likert scale (a 1-to-5 scoring system)*, with a mapping that maps a score range to a specific response (see Table 1). The prompt given to the LLM asks it to generate a score between 0% and 100% on how appropriate the generated label is for the given relationship, and the enumeration then maps that score into one of the response categories and returns the related evaluation category as the LLM's response.

[6] https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

Table 1
Mapping the score ranges with their associated (Likert scale) responses

Score Range	Response
0% - 20%	No
20% - 40%	Unlikely
40% - 60%	Possible
60% - 80%	Likely
80% - 100%	Yes

Finally, as the temperature hyperparameter can be used to control the "level of creativity" found in the model's output [27], several experiments were conducted using the following different temperature settings for the Gemini-flash-2.0 model when assessing the labels themselves: 0.3, 0.5, 0.7, and 1.0.

3.2. In-context Learning

In addition to differing context levels, we also investigated the effect of in-context learning on the response of the LLM. In-context learning refers to the prompt engineering practice of providing the LLM with example prompts and responses so that the LLM generates a response in-line with the examples [28]. We provide the LLM with two levels of in-context learning with each of the different context prompts: zero-shot, where no examples were provided; and few-shot, where the LLM was provided with the five examples shown in Table 2.

Table 2
Examples provided to the LLM for in-context (few-shot) learning.

Domain Label	Range Label	Relationship Label	Evaluation
Sample	Barcode	hasBarcode	Yes
derived	xsd:boolean	isDerivedSample	No
Method	name	hasOptionalMethodName	Likely
Method	Author	wasPerformedBy	Possible
id	xsd:ID	UniqueIdentifier	Unlikely

3.3. Dataset

To assess the ability of the LLM (using Gemini-flash-2.0) to evaluate the appropriateness of the generated labels, we created a dataset of thirty randomly selected relationships present in the AnIML-core schema, with labels that were generated (using ChatGPT-4o) as discussed in Section 3.1. A domain expert (DE) familiar with the domain was recruited and provided with these relationships and their labels. They were then asked to evaluate the labels with respect to how accurately they represented the nature of their understanding of the relationship, using the Likert-scale categories defined in Table 1.

The dataset employed in the empirical study is shown in Table 3. We compare the DE's evaluations against those generated by the LLM. If they are an exact match, we score that evaluation with an accuracy of 1.0. If the evaluation falls into one of the possible two adjacent categories, such as "Likely" when the DE's evaluation was "Yes", we score that evaluation with an accuracy of 0.5. For all other evaluations we score an accuracy of 0.

4. Results

In the evaluation (discussed in Section 3), three experimental parameters were described: *level of context*, *level of in-context learning*, and *temperature*. These parameters were modified to observe the effect they have on the quality of the responses generated by the LLM. The mean scores for each of these evaluations are given in Table 4, and the corresponding variance for each result appears in Table 5. Though our tests, we identify three parameter settings that produce the highest quality evaluations:

Table 3
Experimental Dataset. The evaluation indicates the evaluation of the Domain Expert (DE) on the correctness of the LLM-generated relationship label.

Domain Label	Range Label	Relationship Label	Evaluation
Author	userType	isAuthoredBy	Unlikely
Device	Name	hasCommmonName	Likely
PlotScaleType	ShortTokenType	Non-Plottable Specification	No
Diff	NewValue	hasNewValue	Yes
EmbeddedXML	xsd:string	XML Schema Datatype Governing	No
role	ShortTokenType	playsRoleIn	No
ExperimentStep	comment	hasComment	Yes
sourceDataLocation	ShortStringType	OriginalDataSourceIdentifier	No
ParentDataPointReferenceSet	ParentDataPointReference	containsParentDataPointReference	Possible
id	xsd:ID	Identifier for digital signature anchor point	No
SampleReference	id	refersToSampleID	Yes
IndividualValueSet	endIndex	hasEndIndex	Yes
EncodedValueSet	endIndex	hasEndIndex	Yes
ExperimentDataBulkReference	dataPurpose	consumesDataForPurpose	Possible
comment	ShortStringType	Descriptive Text	Unlikely
AutoIncrementedValueSet	startIndex	hasStartIndex	Yes
experimentStepID	shortTokenType	identifies	No
id	xsd:ID	Identifier Type	Possible
Reason	xsd:string	ReasonExplaination	No
Tag	value	hasValueOrReference	Yes
name	ShortTokenType	Plain-text Data Type	Possible
sample	locationInContainer	hasLocationInContainer	Yes
Author	Affiliation	hasAffiliation	Yes
SeriesSet	name	hasName	Yes
role	ShortTokenType	hasDataType	Yes
AuditTrailEntry	Reference	references	Yes
SampleReferenceSet	SampleInheritance	hasSampleInheritance	Yes
SignatureSet	Signature	containsDigitalSignature	Yes
SeriesSet	Series	containsSeries	Yes
Series	dependency	hasDependencyStatus	Yes

Test A: a Domain and use case prompt using few-shot learning at a temperature of 0.3;
Test B: a Documentation prompt using few-shot learning at temperatures of 0.7; and
Test C: a Documentation prompt using few-shot learning at temperatures of 1.0.

These results appear to fit our hypothesis stated in Section 3.1, that providing increasing context to an LLM surrounding a task will result in increased quality of responses. Whilst this general trend is seen in the results of Table 4, the results for Test A run contrary to this trend, as although the performance is strong, there is a corresponding decrease in accuracy by 0.03 as the amount of provided context is increased by including documentation at a temperature of 0.3. Due to the nature of LLMs being black boxes, we cannot provide a definitive reason as to why the LLM performs better in this specific case. However, given that the temperature parameter controls the "creativity" of the LLM [21], it is possible that a more concise prompt reduces the need for abstract reasoning from the LLM, and the provision of examples aids the LLM to better follow the given prompt. The converse of this statement appears to apply with the documentation prompt (which includes more lexically dense content) which performs better when used with a higher temperature, thus allowing for more creativity. In order to test this further, we took the prompts from Tests A and C, and removed the definition of a relationship from these prompts, as it had been noted that when few-shot learning was used, the addition of the definition of a relationship resulted in a reduction in accuracy. Repeating these tests with the modified prompts confirmed this, as we observed an increase in accuracy for Test A to 0.52, but a drop in accuracy for Test C to 0.47; these results appear to fall in line with our earlier findings.

With the aim of using an LLM as a knowledge evaluation agent as part of a greater ontology engineering pipeline, the consistency of the evaluations is an additionally important metric to consider.

Table 4
Mean accuracy when compared against the evaluation of a DE. *Base* refers to the base prompt defined in Section 3.1

	Zero-Shot				Few-Shot			
	0.3	0.5	0.7	1.0	0.3	0.5	0.7	1.0
Base	0.33	0.32	0.27	0.32	0.42	0.43	0.43	0.37
Definition	0.38	0.37	0.35	0.35	0.32	0.30	0.33	0.35
Domain and use case	0.33	0.35	0.33	0.32	**0.5**	0.45	0.4	0.45
Documentation	0.4	0.37	0.37	0.42	0.47	0.48	**0.5**	**0.5**

Table 5
The variance of the scores when compared against the evaluation of a DE. *Base* refers to the base prompt defined in Section 3.1

	Zero-Shot				Few-Shot			
	0.3	0.5	0.7	1.0	0.3	0.5	0.7	1.0
Base	0.09	0.08	0.08	**0.06**	0.12	0.13	0.15	0.15
Definition	0.10	0.10	0.09	0.09	0.09	0.10	0.11	0.16
Domain and use case	0.09	0.11	0.09	0.09	0.14	0.16	0.16	0.11
Documentation	0.09	0.10	0.10	0.09	0.12	0.13	0.17	0.14

To evaluate the consistency of the evaluations, we show the variance of the accuracies from each experiment in Table 5. The result with the lowest variance was observed when using the base prompt and zero-shot learning at a temperature of 1.0 (indicated in bold in Table 5). However the accuracy of this prompt is only 0.32, suggesting that the quality of the responses, whilst consistent, are poor. Comparing the variances of the three best experiments, we observe that Test A and Test C have the joint lowest variances at 0.14, making these prompts strong candidates for use in a greater pipeline. Whilst the accuracies recorded in Table 4 may initially appear to be low (i.e. with a value of 0.5 or lower), these results are still promising. As described in Section 3.3, an evaluation that falls into a Likert-scale category adjacent to the expected category (for example, if "likely" is generated whilst "yes" is expected), it is given an accuracy of 0.5. Whilst this evaluation is not perfect, in many cases it is still an acceptable evaluation of the relationship. This entails that an accuracy of 0.5 means that the average evaluation generated would be considered acceptable.

5. Conclusion

In this work, we investigated three research questions (Section 1) that explored the viability of using an LLM as a proxy agent for a Domain Expert in evaluating the correctness of a generated relationship label as part of a proposed ontology. The first question, RQ 1, was answered by finding that whilst an LLM may struggle to exactly match the evaluation of a relationship label done by a DE, given the correct prompt format, examples, and LLM parameter settings, the average generated evaluation could be considered accurate. We also found that in most cases, providing an LLM with more context surrounding a task leads to higher quality results. However, we also note that this is not always the case, as in some cases such as where the temperature of the LLM is set to a lower value, an LLM may generate higher quality responses from more concise prompts. This finding addressed RQ 2.

Throughout our experiments, several issues were identified that were observed when using LLMs to evaluate the quality of labels. A primary issue is the lack of consistency (i.e. the stochasticity) of LLMs. It was observed that if the same prompt is provided to an LLM twice, in some cases, two unique responses were generated. This can be an issue when employing the LLM as a knowledge evaluation agent as it is important to know that relationships are being evaluated consistently to be able to reduce the amount of human validation required. Another major issue of employing LLMs in the ontology engineering process is the lack of transparency in LLMs. As they are considered black boxes, their inner workings and biases are unknown. This makes prompt engineering more of an art than a science

as there is no way to know if an optimal prompt has been achieved. To mitigate this issue, we aim to reduce the complexity of the tasks carried out by the LLM, allowing for simple modifications to prompts to have a clear impact on the generated results. The identification of these issues addresses RQ3.

A notable limitation of this work is the fact that a single DE created the ground truth evaluations. This potentially introduces bias, as the way that this DE interacts with and understands the schema shapes their evaluation of generated labels for a given relationship. To mitigate against this bias in future work, we aim to involve a team of DEs that will evaluate the relationships independently, and then come together to produce an agreed upon set of ground truth evaluations. An alternate method to mitigate this bias would be to introduce the original schema author as a DE, as they would have a clearer understanding of each intended relationship. Thus, the author's evaluation of the LLM generated labels for these relationships would be as close to objective as possible.

Overall we conclude that LLMs have potential to be used as knowledge evaluation agents. Whilst the evaluations of LLMs do not perfectly align with those of DEs, we found that the quality of the evaluations can be satisfactory in the majority of cases. This is a valuable finding as it can justify the use of LLMs as knowledge evaluation agents in ontology engineering pipelines, thereby reducing the amount of human time and effort required to validate the synthetic knowledge injected into the ontology by other LLMs.

Acknowledgements

This work has been funded by an EPSRC ICASE studentship, 201146 with Unilever PLC.

Declaration on Generative AI

The author(s) have employed Generative AI tools as part of the contribution and evaluation, but Generative AI tools were not employed to alter or improve the text.

References

[1] IDC, Statista, Various sources, Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028 (in zettabytes) [graph], https://www.statista.com/statistics/871513/worldwide-data-created/, 2024. [Accessed 21-03-25].

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, number 22 in Synthesis Lectures on Data, Semantics, and Knowledge, Springer, 2021. URL: https://kgbook.org/. doi:10.2200/S01125ED1V01Y202109DSK022.

[3] S. Bechhofer, F. Van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, et al., Owl web ontology language reference, W3C recommendation 10 (2004) 1–53.

[4] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, Artificial Intelligence 302 (2022) 103627.

[5] G. Hannah, T. R. Payne, V. Tamma, A. Mitchell, E. Piercy, B. Konev, Towards a methodology for the semi-automatic generation of scientific knowledge graphs from xml documents, in: OM-2023: The 18th International Workshop on Ontology Matching CEUR Workshop Proceedings, volume 3591, CEUR-WS. org, 2023, pp. 85–90.

[6] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[7] M. Hacherouf, S. N. Bahloul, C. Cruz, Transforming xml documents to owl ontologies: A survey, Journal of Information Science 41 (2015) 242–259.

[8] H. Bohring, S. Auer, Mapping XML to OWL ontologies, in: K. P. Jantke, K. Fähnrich, W. S. Wittig (Eds.), Marktplatz Internet: von E-Learning bis E-Payment, 13. Leipziger Informatik-Tage, LIT 2005, 21.-23. September 2005, Leipzig, volume P-72 of *LNI*, GI, 2005, pp. 147–156. URL: https://dl.gi.de/handle/20.500.12116/24882.

[9] M. Ferdinand, C. Zirpins, D. Trastour, Lifting xml schema to owl, in: International conference on web engineering, Springer, 2004, pp. 354–358.

[10] J. F. Sequeda, D. P. Miranker, A pay-as-you-go methodology for ontology-based data access, IEEE Internet Computing 21 (2017) 92–96.

[11] OpenAI, Introducing ChatGPT, https://openai.com/index/chatgpt/, 2022. [Accessed 24-03-2025].

[12] R. Alharbi, V. Tamma, F. Grasso, T. R. Payne, The role of Generative AI in competency question retrofitting, in: Extended Semantic Web Conference, ESWC2024, Hersonissos, Greece, 2024.

[13] R. Alharbi, V. Tamma, F. Grasso, T. R. Payne, A review and comparison of competency question engineering approaches, in: Knowledge Engineering and Knowledge Management, Springer Nature Switzerland, Cham, 2025, pp. 271–290.

[14] A. S. Lippolis, M. Ceriani, S. Zuppiroli, A. G. Nuzzolese, Ontogenia: Ontology Generation with Metacognitive Prompting in Large Language Models, in: Poster and demos track, Satellite proceedings of ESWC2024, 2024.

[15] M. J. Saeedizade, E. Blomqvist, Navigating ontology development with large language models, in: European Semantic Web Conference, Springer, 2024, pp. 143–161.

[16] Y. He, J. Chen, H. Dong, I. Horrocks, Exploring large language models for ontology alignment, 2023. URL: https://arxiv.org/abs/2309.07172. arXiv:2309.07172.

[17] S. Hertling, H. Paulheim, Olala: Ontology matching with large language models, in: Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP '23, ACM, 2023. doi:10.1145/3587259.3627571.

[18] Z. Qiang, W. Wang, K. Taylor, Agent-om: Leveraging llm agents for ontology matching, 2024. URL: https://arxiv.org/abs/2312.00326. arXiv:2312.00326.

[19] L. Yao, J. Peng, C. Mao, Y. Luo, Exploring large language models for knowledge graph completion, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.

[20] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: International Semantic Web Conference, Springer, 2023, pp. 408–427.

[21] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM computing surveys 55 (2023) 1–38.

[22] R. Barile, C. d'Amato, N. Fanizzi, Lp-dixit: Evaluating explanations for link prediction on knowledge graphs using large language models, in: THE WEB CONFERENCE 2025, 2025.

[23] S. Shankar, J. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, I. Arawjo, Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 2024, pp. 1–14.

[24] M. Desmond, Z. Ashktorab, Q. Pan, C. Dugan, J. M. Johnson, Evalullm: Llm assisted evaluation of generative outputs, in: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024, pp. 30–32.

[25] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, R. A. Metoyer, Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, 2025, pp. 952–966.

[26] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, Llms-as-judges: a comprehensive survey on llm-based evaluation methods, arXiv preprint arXiv:2412.05579 (2024).

[27] P.-H. Wang, S.-I. Hsieh, S.-C. Chang, Y.-T. Chen, J.-Y. Pan, W. Wei, D.-C. Juan, Contextual temperature for language modeling, arXiv preprint arXiv:2012.13575 (2020).

[28] G. Hannah, R. T. Sousa, I. Dasoulas, C. d'Amato, On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs, Journal of Web Semantics 84 (2025) 100843.